To appear in the Proc. of the European Conf. on Computer Vision, 1998.

Robust Video Mosaicing through Topology Inference and Local to Global Alignment

Harpreet S. Sawhney, Steve Hsu, and R. Kumar

Vision Technologies Laboratory
Sarnoff Corporation
CN5300, Princeton, NJ 08543, USA
{hsawhney,shsu,rkumar}@sarnoff.com

Abstract. The problem of piecing together individual frames in a video sequence to create seamless panoramas (*video mosaics*) has attracted increasing attention in recent times. One challenge in this domain has been to rapidly and *automatically* create high quality seamless mosaics using inexpensive cameras and relatively free hand motions.

In order to capture a wide angle scene using a video sequence of relatively narrow angle views, the scene needs to be scanned in a 2D pattern. This is like painting a canvas on a 2D manifold with the video frames using multiple connected 1D brush strokes. An important issue that needs to be addressed in this context is that of aligning frames that have been captured using a 2D scanning of the scene rather than a 1D scan as is commonly done in many existing mosaicing systems.

In this paper we present an end-to-end solution to the problem of video mosaicing when the transformations between frames may be modeled as parametric. We provide solutions to two key problems: (i) automatic inference of topology of the video frames on a 2D manifold, and (ii) globally consistent estimation of alignment parameters that map each frame to a consistent mosaic coordinate system. Our method iterates among automatic topology determination, local alignment, and globally consistent parameter estimation to produce a coherent mosaic from a video sequence, regardless of the camera's scan path over the scene. While this framework is developed independent of the specific alignment model, we illustrate the approach by constructing planar and spherical mosaics from real videos.

1 Motivation

Creation of panoramic images and mosaics from a video sequence or a collection of images has attracted tremendous attention from researchers and commercial practitioners alike. However, most previously developed systems for creating panoramas have either been limited by their use of special fixtures (e.g. tripods) for precisely controlled image capture [3] or have been restricted to a

one-dimensional [1] scanning of a scene [8]. In this paper we present some key technical advances in automatically creating mosaics from video sequences that users create by smoothly moving a hand-held camera to cover the whole scene using multiple overlapping swaths (or 2D scanning). The overlap between consecutive swaths may be quite small relative to the width or height of individual images. Significant parts of this technology have been packaged into consumer level applications for various kinds of mosaicing scenarios under the commercial name of *VideoBrush*^{TM} [15].

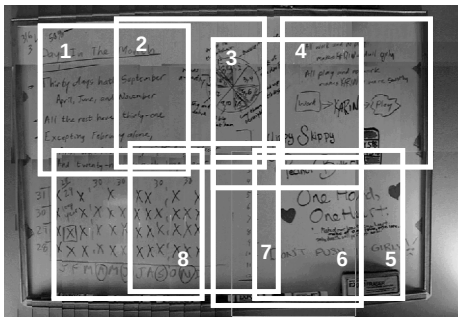

Fig. 1. An S-pattern for scanning a planar object.

Unlike the 1D case, in 2D scanning, there is a wrap-around and frames belonging to neighboring swaths must align with each other even though they are not temporally adjacent frames. Our framework is primarily motivated by the observation that although alignment of consecutive frames may be accurate, simply concatenating local alignment models leads to gross global misalignments. For example, in Fig. 1, if the capturing order of frames is $1-8$, and each frame is aligned with its predecessor, there is no guarantee that frames 1 and 8 will be aligned when warped to a global reference coordinate system.

There are two key problems in video mosaicing. First is to automatically infer the 2D neighborhood relations (topology) among frames. The input video sequence just provides a time ordered collection of frames that have captured overlapping parts of the scene. To create a mosaic, these frames need to be placed in their correct 2D topology. Second, the frames need to be stitched together so that the scene is represented as a single seamless whole even when the overlaps between frames may vary widely over the complete 2D scan. Previous work [4, 10, 13, 16] has addressed only a subset of the issues involved with alignment of frames but has not provided a comprehensive framework within which automatic inference of 2D topology is combined with local-to-global alignment for an end-to-end solution to the mosaicing problem.

The framework that we present has been applied to a number of different applications involving capture of a wide angle scene from multiple narrow angle views. In this work we will illustrate the framework through its applications to scanning planar objects with unrestricted camera motion, and creation of (hemi-)spherical mosaics by systematically scanning a scene from an approximately fixed position. Each of these applications highlights a different parameterization of the alignment model that is required for mosaicing. However, the framework within which the different parameterizations work is the same.

[1] We use the term 1D scanning for an arrangement of frames in which each frame may overlap only with its two immediate temporal neighbors. In contrast 2D scanning of a scene allows frames to overlap even when they may not be temporal neighbors.

2 Issues and Related Work

For a video sequence of a planar scene, or of a 3D scene captured with approximately a fixed camera center, the process of creating a seamless mosaic of the scene involves the following major issues:

1. Determination of the 2D topology of the sequence of frames on a suitable 2D manifold.
2. Parameterization and estimation of the transformations that can be used to establish the geometric relationships between the various frames and a global mosaic reference frame.
3. Handling lens distortion.
4. Creation of a seamless mosaiced representation of the scene so that new views may be created.

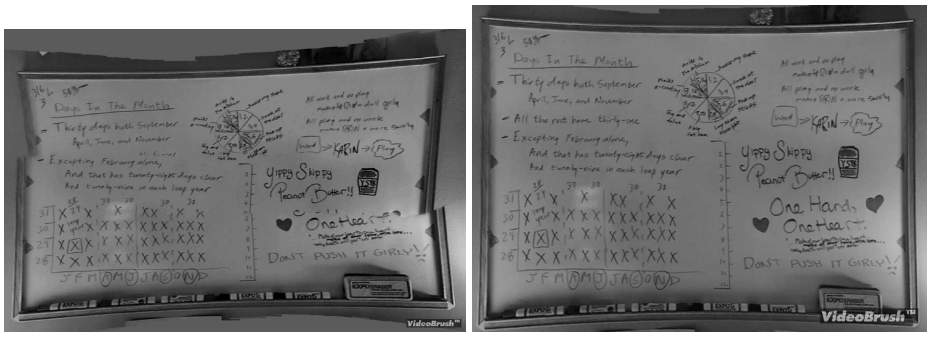

Fig. 2. Left: 75 frames captured using 3 successive horizontal swipes are poorly mosaiced using frame-to-frame plane projective parameters. **Right**: 75 frames captured using 3 successive horizontal swipes are successfully mosaiced using frame-to-mosaic plane projective parameters.

Previous approaches to the problem of video mosaicing have not addressed all these issues comprehensively or robustly. In particular, the important issue of automatic topology determination has been completely ignored. Early works on mosaics and image alignment like [4, 6, 9, 12, 14] essentially did 1D/2D mosaicing by solving for the alignment parameters of each frame individually. Each frame is registered to the previous frame in the sequence. The frame-to-frame registration parameters are concatenated to compute the global frame to mosaic alignment parameters. A mosaic is subsequently created by warping each frame to a reference coordinate system using the computed alignment parameters. The main problem with this approach is that the parameters obtained by the alignment of frames locally are not accurate enough to assemble a well aligned mosaic. Small errors in computing the frame-to-frame alignment parameters, or small deviations of the model from the true model, get compounded when they are concatenated to compute the global frame to mosaic alignment parameters.

An example of such a case is shown in Fig. 2 (left). A video sequence of 75 frames is captured of a planar white-board. Between each pair of consecutive frames, an 8-parameter projective transformation is solved for using a multi-resolution direct alignment technique [10, 12]. Finally, the center frame is chosen as the reference frame and each of the other frames is warped using the relevant concatenated pairwise parameters. It is to be emphasized that in spite of each of the pairwise frame alignments being almost perfect, the resulting mosaic suffers from gross misalignments. Therefore, one important issue to be addressed is how multiple frames can be constrained to create parameters that align all the frames accurately.

An improvement over frame to frame registration is to align each new frame to the mosaic as it is being built [10]. This provides a larger image context for aligning each new frame, and captures some constraints from spatially neighboring frames that are not temporally adjacent. Fig. 2 (right) shows the mosaic built using the frame-to-mosaic alignment algorithm. Unlike the frame-to-frame alignment algorithm (Fig. 2 (left)), the frame-to-mosaic algorithm produces a perfect mosaic. The frame-to-mosaic algorithm works well as long as there is sufficient overlap between frames in neighboring swaths, for instance in the top left-to-right and bottom right-to-left swaths in Fig. 1. However, when there is very little overlap, then the performance of the frame-to-mosaic algorithm is very similar to the frame-to-frame algorithm. Such an example can be seen in Fig. 3 (left). The frame-to-mosaic algorithm produces a poor alignment for one of frames in the bottom swath. This one poor alignment causes an overall error in the mosaic construction. The main problem with the frame-to-mosaic method is that computing the alignment is still a causal 1D processing chain. As each frame is registered, its pose with respect to the mosaic is committed with no possibility of improvement based on frames that are processed subsequently. If any link on this chain is bad, the overall mosaic construction is erroneous.

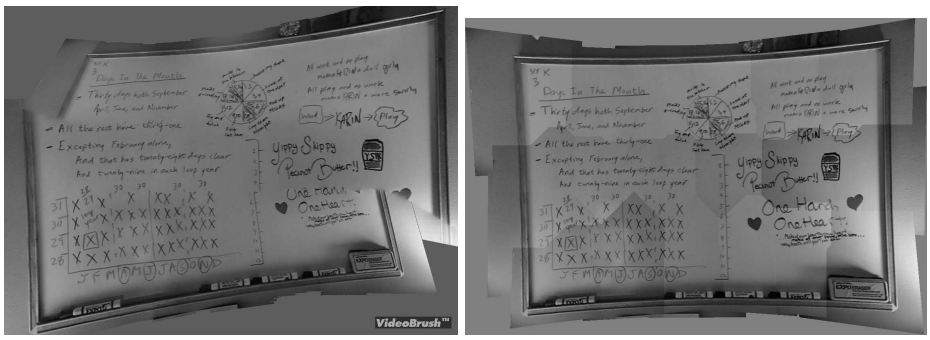

Fig. 3. Left:75 frames captured using 3 successive horizontal swipes are poorly mosaiced using frame-to-mosaic plane projective parameters.
Right:75 frames captured using 3 successive horizontal swipes are succesfully mosaiced using local-to-global plane projective parameters.

Some work in the area of image-based rendering has explicitly concentrated on creating closed mosaics on 1D manifolds for instance by projecting on cylindrical surfaces and then developing the surface [3, 7]. It has been explicitly assumed that the collection of images captures the complete 360^o 1D manifold. This constraint has been enforced by using tripods and systematic angular steps between frames. In [5], for the 1D case, closure of the 1D mosaic is enforced by adjusting the focal length so that a complete 360^o coverage is obtained. Therefore, in these works, it is assumed a priori that the collection of frames when mapped on a 1D manifold closes on itself.

Recently the problem of stitching together video frames on a 2D manifold has attracted attention. In a recent paper [13], Szeliski and Shum parameterize the 2D mosaicing problem using rotations and focal length of the camera that is moved around a fixed point. However, the various issues listed earlier have not been comprehensively addressed. The 2D topology was specified manually in their approach.

The determination of 2D topology or the relative placement of frames is critical in achieving simultaneous alignment of all the frames to create a seamless representation. In this paper, we combine automatic inference of 2D topology with a local to global alignment strategy. In our technique, each frame is able to constrain all other frames through constraints determined by the topology. For example, Fig. 3 (right) shows the perfect mosaic construction from the white-board video using the new algorithm, in contrast to the frame-to-mosaic alignment method (Fig. 3 (left)).

3 Our Approach

We provide a framework for the creation of mosaiced representations of a planar scene, or of a 3D scene from approximately a fixed viewpoint, by providing solutions to and integrating means for 2D topology determination, local alignment, and global consistency. Planar topology is adequate when the captured sequence can be represented as a whole on a planar surface. This is the case when the goal is to produce high resolution planar mosaics of paintings, white-boards and 3D scenes from relatively small total viewing angles. However, when the scene is captured through a sequence that closes on itself or covers the complete sphere or a significant part of a sphere around the fixed viewpoint, then the planar topology is not adequate for seamless representation of the scene since plane projective mappings will map points to infinity.

A 2D manifold, e.g. a plane or a sphere, may be explicitly used for representing planar images on the manifold. Alternatively, the specific transformations used to map points between frames can be implicitly used to represent the input frames on a manifold. In either case, 2D topology determination figures out which frames overlap and hence are neighbors on the appropriate manifold. The topology determination is an iterative process since it can be inferred only when frames are placed with respect to each other by solving for specific transformations. A hypothesis for a new neighborhood relationship is tested for reliability using a quality measure that is computed after coarse and fine alignment of the neighbors.

Local coarse registration estimates a low complexity (say, 2D translational only) mapping between neighbors. Fine local registration estimates a higher complexity mapping between neighbors or between a frame and the current estimate of the local mosaic. The step of globally consistent optimization infers all the reference-to-frame mappings by simultaneously optimizing them, such that they are maximally consistent with all the local registration information. Global consistency may be imposed by solving for purely parametric alignment models corresponding to the 2D manifold over which the mosaic is represented. This is similar to the bundle block adjustment used in photogrammetry [11]. In addition, misalignments due to departures from the global models are handled by quasi-parametric or piecewise parametric alignment between regions of overlap. However, in this paper we will not dwell on quasi-parametric alignment. For accurate alignment, we iterate between the coarse matching, topology determination and global consistency steps.

The complete process is depicted in Fig. 4.

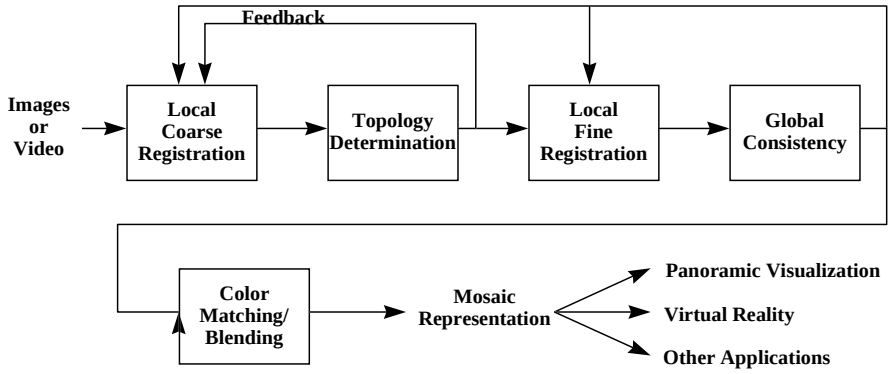

Fig. 4. Block diagram of the main steps.

While the previously described stages accomplish geometric alignment of the source images, color matching/blending adjusts for discrepancies in color and brightness between images. This is critical for avoiding seams in the mosaic. Another important real-world issue that needs to be taken care of especially for consumer cameras is that of handling lens distortion. We estimation lens distortion within the framework of multi-frame alignment in which we can solve for the alignment and lens distortion parameters without the need of a calibration object. The computed lens distortion parameter is applied as a correction to all the frames before mosaic creation. The reader is referred to [10] for details. We will not be dwelling on blending and lens distortion correction in this paper, but it is to be emphasized that in a real application, solutions to both are a must for creation of high quality mosaics on 2D manifolds.

4 Problem Formulation

The models that relate multiple frames when a camera is capturing a planar scene under arbitrary motion, or a 3D scene from a fixed viewpoint, can be parameterized as plane projective, or 3D rotations, or 3D rotations and translations with the 3D plane parameters. In order to create a seamless mosaic of a collection of frames, an optimal set of reference-to-image mappings need to be created: $\mathbf{u} = \mathbf{P_i}(\mathbf{x})$ of a parametric class, where $\mathbf{x}$ denotes a point on the reference surface and $\mathbf{u}$ is a point on the ith source image. In general, the shape of the reference surface and the source image surfaces can be any 2D manifold in 3D, such as planes and spheres.

Denote the mosaic image as $M(\mathbf{x})$, and the source images by $I_i(\mathbf{u})$. The $\mathbf{P_i}$'s need to be computed so that for each $\mathbf{x}$, the point $\mathbf{P_i}(\mathbf{x})$ in every image i corresponds to the same point in the physical scene. This condition assures that the mosaic image $M(\mathbf{x})$ constructed by combining pixels $\{I_i(\mathbf{P_i}(\mathbf{x})), \forall i\}$ will yield a spatially coherent mosaic, where each point $\mathbf{x}$ is in one-to-one correspondence with a point in the scene.

We formulate the problem of optimal mosaic creation as that of minimizing an objective function that measures the misalignment between frames as well as the redundancy in information. Formally, the following function is optimized:

$$\min_{\{\mathbf{P_i}\}} \sum_{\mathbf{x}} \mathrm{var}_i\{I_i(\mathbf{P_i}(\mathbf{x}))\} + \sigma^2(\text{Area of the mosaic}) \tag{1}$$

where σ is a scale factor and $\mathrm{var}_i\{\cdot\}$ denotes the variance of the pixels from different frames that map to each $\mathbf{x}$. In words, the above Minimum Description Length cost function measures the compactness of representing a collection of frames in the form of a mosaic plus residuals of frames w.r.t. the mosaic. Note that the variances could be measured not just on the intensities directly, but alternatively on filtered representations of image intensities, or on the point locations $\mathbf{P_i}(\mathbf{x})$'s directly.

In order to optimize the error function of Equation (1), we need to maximize the overlap between warped frames in the mosaic coordinate system by finding the globally optimal alignment parameters. The optimization in general cannot be done in a closed-form. Therefore, we adopt an iterative technique. The technique is based on the observation that if the 2D topology of the input frames is known on an appropriate 2D manifold, and the local alignment parameters (or correspondences) are available between neighboring frames, then global bundle block adjustment can be used to solve for accurate $\mathbf{P_i}$'s. On the other hand if approximate knowledge of $\mathbf{P_i}$'s is available, then neighborhood relations can be inferred that can further establish new relationships between frames. Our approach switches between the two steps of topology determination and parameter estimation iteratively to reach a globally optimal solution. The topology determination step hypothesizes local neighborhood relations, and the global optimization step uses the local constraints to determine the parameters. In between these two steps, we need to establish correspondence relations between the neighboring frames and verify these with a quality measure. This is achieved using local coarse and fine alignment (see Fig. 4).

4.1 Topology Determination

The 2D topology refers to the set of pairs of input frames (i, j) whose local alignment information participates in globally consistent parameter estimation. Such pairs, henceforth called neighbors, are necessarily images with sufficient spatial overlap to make local estimation feasible.

At the start of the first iteration, there is typically no information on the $\mathbf{P_i}$'s whatsoever; hence, under the reasonable assumption that consecutive frames of a video sequence are overlapping, the initial topology defaults to a linear chain of temporal neighbors. Local alignment of such neighbors and global consistency—a trivial concatenation of motion models—yield the first estimate of $\mathbf{P_i}$.

In subsequent iterations, topology determination may become nontrivial and essential. Non-consecutive frames may be discovered to be neighbors, such as frames in adjacent swipes of an S pattern or pairs which close a loop or spiral scan. These patterns can be formed on any shape of reference surface if the direction of camera motion changes. In case of a closed shape like a sphere, moreover, loops can be formed even with constant camera motion, as typified by scanning a 360° panorama. Because topology is inferred from only approximate knowledge of $\mathbf{P_i}$ and because the choice of surface shape may be changed during the course of global consistency (e.g. from planar to spherical), it is possible that not all proper neighbors will be found during the second iteration; multiple iterations may be required to converge to agreement between topology and parameter estimation.

We update the topology at the start of each iteration by generating hypotheses for new neighbors, which get verified or refuted by local registration, and adding only verified neighbors as arcs in the neighbor relation graph, G. New candidates might be selected using various criteria, including influence on subsequent global estimation and proximity of the images.

1. The existing topology dictates where adding new arcs would have the most effect on the accuracy of the global parameter estimate. The first arc that closes a loop or ties together two swipes is significant, but not one which parallels many other nearby arcs. It is not essential to include every possible overlapping pair in the topology for accurate global alignment, nor is it computationally efficient. Therefore, it is desirable to limit the density of arcs within any local region.
2. The current topology and set of global parameter estimates $\mathbf{P_i}$ determine the relative spatial locations and uncertainty for any pair of frames under consideration. It is desirable to choose pairs which are most likely to overlap and to have least positional uncertainty so that local alignment need not search a large range.

The two desiderata are generally in direct conflict, since arcs of high payoff (influence) are very often image pairs with high risk (positional uncertainty). Our current experiments prioritize candidate arcs by greatest overlap rather than greatest influence, and additionally skip arcs too close to existing ones. It is expected that as iterations progress, global parameter estimates will increase

in accuracy, drawing the high leverage pairs closer until they have reasonable overlap and uncertainty to get registered and added to G.

Specifically, candidates are added by considering their arc length d_{ij} in relation to path length D_{ij}. Arc length is defined by the distance between warped image centers $\mathbf{x_i}$, $\mathbf{x_j}$ on the mosaic surface, normalized by the warped frame "diameter" r_i, r_j:

$$d_{ij} = \frac{\max(0, |\mathbf{x_i} - \mathbf{x_j}| - |r_i - r_j|/2)}{\min(r_i, r_j)}$$

Path length D_{ij} is defined as the sum of arc lengths along the minimum sum path between nodes i, j in the existing graph. To add an arc, d_{ij} must not exceed a maximum limit and must be significantly shorter than D_{ij}, and the image reliability measure ρ_{ij} (see below) must be high. This heuristic tends to select arcs that both have good overlap and will add non-redundant constraints to the global bundle block adjustment.

4.2 Local Coarse and Fine Registration

Local alignment verifies the goodness of the neighborhood relationships suggested by topology determination by aligning images pairwise using parametric models. For each neighboring pair of images, the mapping $\mathbf{Q_{ij}}$ is estimated such that, for each point $\mathbf{u}$ in image i, the point $\mathbf{u}' = \mathbf{Q_{ij}}(\mathbf{u})$ in image j corresponds to the same point in the physical scene. Since the absolute pixel-to-scene calibration is typically not known in advance, this correspondence must be inferred by matching the appearance of $I_j(\mathbf{u}')$ and $I_i(\mathbf{u})$.

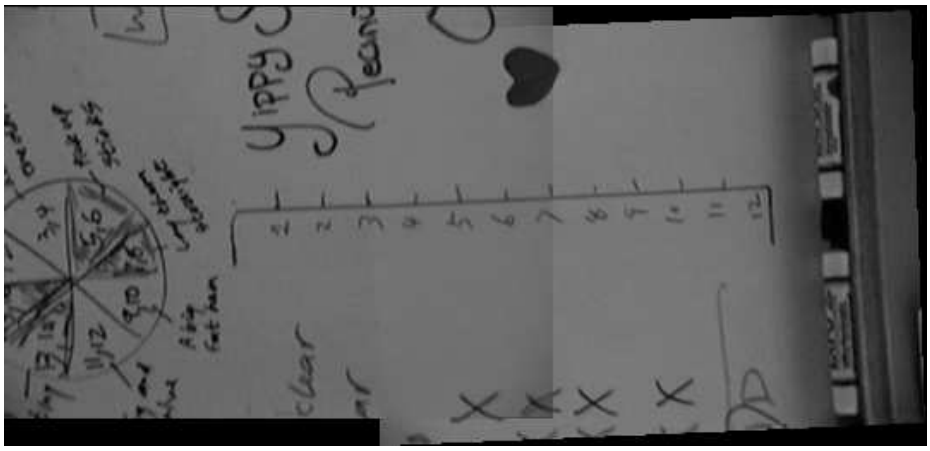

Fig. 5. Local plane projective registration between two neighboring frames of the *Whiteboard* video showing accurate automatic alignment even with small overlap. A mosaic of the two frames is shown. (The mosaic is rotated 90^o to fit on the page.)

Often, $\mathbf{Q_{ij}}$ is restricted to a parametric family of mappings, such as projective mapping or pure translation. While a simple parametric model may only be an approximation of the true mapping, it is often faster and more reliable to estimate than a higher order model. Indeed, the kind of mapping does not even have to be the same as the mapping $\mathbf{P_i}$ to be estimated during global consistency.

In general, the magnitude of motion between temporally contiguous frames may be tens of pixels, and that between other neighbors may be as high as hundreds of pixels. We divide the local alignment problem into steps in which models of increasing complexity are estimated while establishing correspondence. Initially, a large range of 2D translations only is searched to establish robust rough correspondence. The image is divided into multiple blocks and each block establishes its correspondence through coarse-to-fine search with normalized correlation as a match measure. Majority consensus between the blocks is used to compute the 2D translation.

Once a reasonable guess of the translation is available, more accurate alignment is performed by fitting progressively complex models and minimizing the sum-of-squared-differences (SSD) error measure in a coarse-to-fine manner over a Laplacian pyramid [1]. At each level of the pyramid, the unknown parameters are solved for by:

$$\min_{\mathbf{Q_{ij}}} \sum_{\mathbf{u}} (I_j(\mathbf{Q_{ij}}(\mathbf{u})) - I_i(\mathbf{u}))^2 \tag{2}$$

The initial 2D translation parameters at the coarsest level, and subsequently the refined parameters from each level are used to warp I_j and the next increment in the parameters is solved using a Levenberg-Marquardt iteration. Local alignment is done progressively using affine and then projective parameters to establish accurate correspondence between neighboring frames. In general, the progressive complexity technique provides good correspondences between frames even when the overlap may be as low as 10% as is shown in an example in Figure 5.

Before accepting the pair ij as neighbors, a reliability measure ρ_{ij} is computed. This measure is thresholded to discard poor estimates, and is also applied as a weight factor during global consistency. Using the computed $\mathbf{Q_{ij}}$, the resulting reference and warped images are compared by one of the following: (i) the mean (or median) absolute or squared pixel value error; (ii) normal flow magnitude; (iii) normalized correlation to compute ρ. We have found that normalized correlation gives the most reliable measure of alignment.

4.3 Global Consistency

The steps of topology determination and pairwise local alignment lead to local maximal overlaps between frames. If the local alignment parameters were globally consistent too then the cost function of Equation (1) would have been optimized. However, in general, the local alignment parameters provide good correspondences between neighboring frames but may still be far from providing consistent alignment parameters for each frame's mapping to a mosaic coordinate system. This was illustrated by the example in Figure 2. In order to optimize the error function of Equation (1), it is assumed that the topology determination and local alignment have achieved a local minimum of the second term, that is

the area term. Now with the overlap between frames fixed, and based on the correspondences in the overlapping areas provided by local alignment, the first term is minimized with respect to the global alignment parameters.

The jointly optimum set of reference-to-image mappings $\mathbf{P_i}$ is computed by minimizing an alternative form for the first term of Equation (1) that measures point correspondence errors instead of image gray value errors.

$$\min_{\{\mathbf{P_i}\}} E = \sum_{ij \in G} E_{ij} + \sum_i E_i.$$

The term $E_{ij}(\mathbf{P_i}, \mathbf{P_j}; \mathbf{Q_{ij}})$ measures the alignment errors between points in neighboring images that, according to local alignment $\mathbf{Q_{ij}}$, are supposed to map to the same point in the mosaic coordinate system.

The regularization term $E_i(\mathbf{P_i})$ allows the inclusion of any kind of *a priori* desirable characteristic for the reference-to-image mappings. For example, for the case of projective mappings, E_i are designed to penalize the amount of distortion frames must undergo when warped to the mosaic. Another source of knowledge may be physical measurements of camera position and orientation. All such criteria can be expressed as functions of $\mathbf{P_i}$, constituting the error term E_i.

$\mathbf{P_i}$ is often restricted to a parameterized family of mappings, in which case the domain of this optimization problem is a finite-dimensional vector. Nevertheless, the global error criterion E is typically a complicated function of the unknown $\mathbf{P_i}$'s and only iterative solution is possible or practical. The optimization process is terminated either after a fixed number of iterations or when the error stops decreasing appreciably. In practice, typically five iterations of the global alignment have been found to be adequate.

The optimization should be initialized with some reasonable starting estimate of the $\mathbf{P_i}$'s. We choose a spanning tree T of the graph of neighbors G, and begin by optimizing $E = \sum_{ij \in T} E_{ij}$. Since there are no loops in subgraph T, it is possible to minimize this error by simply requiring $\mathbf{P_j}(\mathbf{x}) = \mathbf{Q_{ij}}(\mathbf{P_i}(\mathbf{x}))$ exactly for every pair of neighbors in T. As a special case, if T is all pairs of temporally adjacent frames, then this is nothing more than concatenating a linear chain of frame-to-frame mappings.

5 Examples

This section provides two example scenarios where different parameterizations are employed. Notation: for a 3D vector $\mathbf{X} = (X_1, X_2, X_3)$, define $\Pi(\mathbf{X}) = (X_1/X_3, X_2/X_3)$ and $\hat{\mathbb{V}}(\mathbf{X}) = \mathbf{X}/|\mathbf{X}|$. For a 2D vector $\mathbf{u} = (u_1, u_2)$, define $\tilde{\mathbf{u}} = (u_1, u_2, 1)$. Also, a homography in $\mathcal{P}^2$ is written as $\mathbf{x} \approx \mathbf{A}\mathbf{X}$.

5.1 Planar Mosaic

In order to create a seamless mosaic of a planar surface from a video sequence acquired by a freely moving camera, the reference-to-image mappings as well as relative image-to-image mappings are well described by projective mappings.

Local coarse registration uses a pure translation mapping, for efficiency, while local fine registration uses projective mapping. Topology is recalculated once,

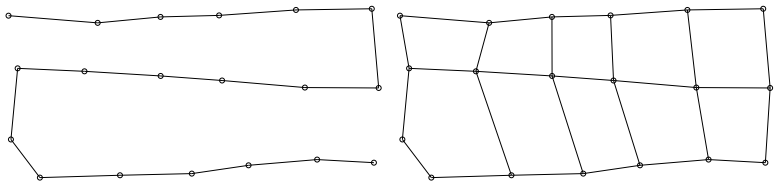

Fig. 6. Topology refinement for an S pattern sequence: (1) Default topology of temporal neighbors only; (2) Topology of spatial neighbors. Topology converged after 2 cycles.

following the local coarse registration, whose translational shift estimates Q_{ij} are simply integrated to give preliminary estimates of P_i.

Global consistency endeavors to determine the jointly optimum reference-to-image mappings of the form $\tilde{\mathbf{u}} \approx A_i^{-1}\tilde{\mathbf{x}}$. The inverse mapping is then $\tilde{\mathbf{x}} \approx A_i\tilde{\mathbf{u}}$.

The complete E consists of two kinds of terms:

1. For each pair of neighboring images,

$$E_{ij} = \sum_{k=1}^{4} \left| \Pi(\mathbf{A_i}\tilde{\mathbf{u_k}}) - \Pi(\mathbf{A_j}\tilde{Q}_{ij}(\mathbf{u_k})) \right|^2$$

   where the $\mathbf{u_k}$ are corners of the overlap between the images (typically four points). This term penalizes inconsistency between reference-to-image mappings and local registration.
2. For each image,

$$E_i = \sum_{k=1}^{2} \left| \Pi(\mathbf{A_i}\tilde{\alpha_k}) - \Pi(\mathbf{A_i}\tilde{\beta_k}) - (\alpha_\mathbf{k} - \beta_\mathbf{k}) \right|^2$$

   where $\alpha_\mathbf{1}$, $\alpha_\mathbf{2}$, $\beta_\mathbf{1}$, $\beta_\mathbf{2}$ are the midpoints of the top, left, bottom, and right sides of the source image. This term penalizes scale, rotation, and distortion of the images when warped to the mosaic. Additionally, the term $|\Pi(\mathbf{A_1}(0,0,1))|^2$ is added to E_1 to fix the translation of one frame.
   In the absence of these terms, the solution for $\{\mathbf{A_i}\}$ is under-determined, since any projective transformation applied to the whole reference coordinate system would not affect E.

The global error is optimized as follows. First, the $\mathbf{A_i}$'s are initialized by concatenating the local registration projective mappings within a spanning tree. Second, the sum of E_i terms only is minimized with respect to the update $\mathbf{A_i} \leftarrow \mathbf{B_0}\mathbf{A_i}$ where $\mathbf{B_0}$ is a common projective mapping. Third, the complete E is minimized with respect to the update $\mathbf{A_i} \leftarrow \mathbf{B_i}\mathbf{A_i}$ where $\mathbf{B_i}$ is a per-image projective mapping. For the last two steps, optimization is done by Gauss-Newton method, which requires only first derivatives of the terms inside $|\cdot|^2$ with respect to the coefficients of $\mathbf{B}$; typically convergence is reached in 3–5 iterations for mosaics of dozens of frames.

The complete topology inference and local-to-global alignment framework for the example of Fig. 3 (right) is illustrated in Figure 6. The first cycle starts with the default topology of temporal neighbors only. The local estimator finds coarse translations Q_{ij}, and global estimation simply concatenates these translations into the reference-to-frame parameters P_i. The second cycle detects non-consecutive spatial neighbors, performs local estimation of projective models, then optimizes the global plane projective parameters. In this instance, topology converges in 2 cycles. The resulting mosaic was shown in Figure 3.

5.2 Spherical Mosaic

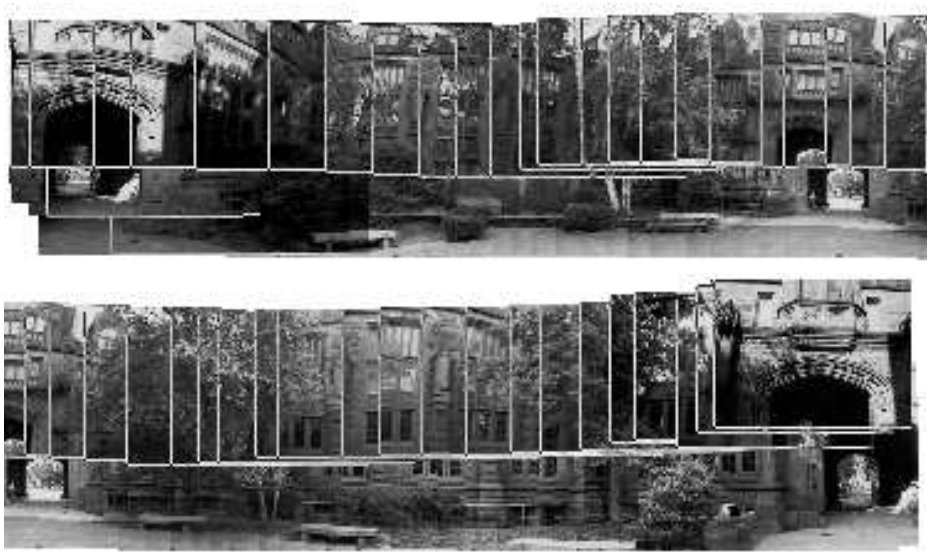

Fig. 7. 184 frames captured using 3 successive horizontal 360° swipes shown as overlapping frames after the initial step of 2D translational alignment between consecutive frames. (The sequence has been broken into two parts for clarity.)

In order to illustrate that our framework for constructing a seamless mosaic representation is general, we now show the creation of seamless mosaics of any 3D scene from a video sequence acquired by a camera rotating about a fixed point. The camera parameters including lens distortion are unknown. In this situation the best shape for the reference surface is a sphere, which places no limitation on the angular extent of the mosaic representation. The image-to-image mappings are still well described by projective mappings, but the sphere-to-image mappings are not. The projective mappings are converted to 3D rotations and camera calibration parameters to infer the 2D topology on a sphere as well as to solve for globally consistent rotations and the calibration parameters.

Local coarse registration uses a 2D rotation/translation mapping, while local fine registration uses projective mapping. Topology is recalculated following the local coarse registration, whose translational shift estimates $\mathbf{Q_{ij}}$ are simply concatenated to give preliminary estimates of $\mathbf{P_i}$.

Global consistency endeavors to determine the jointly optimum reference-to-image mappings of the form $\mathbf{u} = \Pi(\mathbf{F}\mathbf{R_i}^{\mathbf{T}}\mathbf{X})$ where $\mathbf{F}$ is a upper triangular camera calibration matrix, $\mathbf{R_i}$ is an orthonormal rotation matrix, and $\mathbf{X}$ is a 3D point on the unit sphere reference surface. The method of [2] is used to estimate a common $\mathbf{F}$ from all the $\mathbf{Q_{ij}}$'s. Using this estimation the inverse mapping can be written as $\mathbf{X} = \mathbf{R_i}\hat{\forall}(\mathbf{F^{-1}}\tilde{\mathbf{u}})$. It is assumed that the same $\mathbf{F}$ is valid for each frame.

The complete E consists solely of inconsistency terms for pairs of images

$$E_{ij} = \sum_{k=1}^{4} \left| \mathbf{R_i}\hat{\forall}(\mathbf{F^{-1}}\tilde{\mathbf{u_k}}) - \mathbf{R_j}\hat{\forall}(\mathbf{F^{-1}}\tilde{\mathbf{Q}}_{\mathbf{ij}}(\mathbf{u_k})) \right|^2 .$$

For the central image in the mosaic, $\mathbf{R_{i_0}}$ is fixed as the identity.

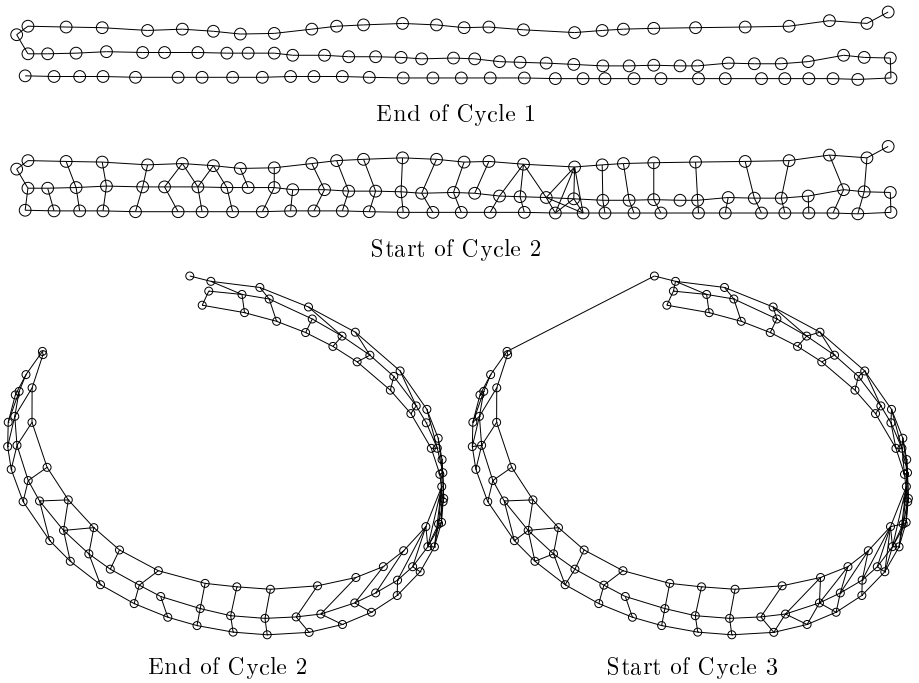

Fig. 8. Topology refinement for a 360° panoramic S pattern sequence: (1) Default topology of temporal neighbors only; (2,start) Planar toplogy of spatial neighbors; (2,end) Spherical topology due to global estimation; (3) Loop closure detected. Topology converged after 3 cycles.

The global error is optimized as follows. First, the $\mathbf{R_i}$'s are initialized by locally minimizing each E_{ij} in a spanning tree. Second, the complete E is minimized with respect to the update $\mathbf{R_i} \leftarrow \mathbf{B_i R_i}$ where $\mathbf{B_i}$ is a per-image rotation matrix, using a Gauss-Newton method. Note, an alternative strategy is to update both the common $\mathbf{F}$ matrix and the individual $\mathbf{R}$ matrices during each iteration of the non-linear optimization of function E.

The complete topology inference and local-to-global alignment framework for a spherical mosaic surface is illustrated in Figure 8 for a video sequence containing three successive horizontal 360° swipes. As a preprocessing step, lens distortion is estimated and compensated using the method of [10]. The first cycle is the same as the planar mosaic case, yielding P_i's which place the images on the reference surface as shown in Figure 7. The second cycle detects non-consecutive spatial neighbors, performs local estimation of projective models, estimates $\mathbf{F}$, then optimizes the global alignment using a spherical parameterization. The angular field of view of each frame implied by $\mathbf{F}$ is $37 \times 29°$. At the end of the second cycle, the 360° panorama does not quite close; however, the ends are near enough so that that during the third cycle the loop closure is hypothesized, verified, and incorporated into the globally consistent estimation. In this instance, topology converges in 3 cycles.

$\mathbf{F}$ estimation depends heavily on the non-affine parameters of the $\mathbf{Q_{ij}}$, which are in turn sensitive to alignment of the periphery of the field of view, namely the image area most affected by lens distortion. If lens distortion had not been compensated, the $\mathbf{F}$ estimated in this example would have implied a $29 \times 23°$ field of view for each frame. This would have made the spherical topology at the end of the second cycle wrap only half-way around the circle, leaving the ends of the topology too distant for loop closure to be hypothesized.

The final seamless mosaic is shown in Figure 9. For hardcopy presentation, a Mercator projection readily portrays the full contents of the panoramic scene. For a compelling visual effect, however, it is preferable to view the spherical mosaic using a 3D graphics system, such as a VRML browser or stereo display.

6 Conclusions and Future Work

The new framework for robust image alignment and mosaic construction iterates among automatic topology determination, local alignment, and globally consistent parameter estimation to produce a coherent mosaic from a video sequence. The framework captures constraints missed by prior methods, namely constraints between non-consecutive but spatially neighboring frames, and non-causal constraints from frames appearing later in time. This approach allows any pattern of scanning of the scene.

We are extending the framework in several directions to improve its flexibility and efficiency. Quasi-parametric and explicit 3D recovery can be incorporated to handle scenes and scanning modes in which the parallax effects are significant. The local to global paradigm will be generalized to a hierarchy of local mosaics of clusters of frames. This multi-scale partitioning of the neighborhood constraints

not only speeds up convergence of very large scale topologies but also allows relaxing the rigidity of parametric models.

References

[1] J. R. Bergen et al. Hierarchical model–based motion estimation. In *Proc. 2nd European Conference on Computer Vision*, pages 237–252, 1992.

[2] R. I. Hartley. Self-calibration from multiple views with a rotating camera. In *ECCV*, pages 471–478, 1994.

[3] Apple Computer Inc. An overview of apple's QuickTime VR technology, 1995. http://quicktime.apple.com/qtvr/qtvrtech5_25.html.

[4] M. Irani, P. Anandan, and S. Hsu. Mosaic based representations of video sequences and their applications. In *Proc. Intl. Conf. on Computer Vision*, pages 605–611, 1995.

[5] S. B. Kang and R. Weiss. Characterization of errors in compositing panoramic images. In *Proc. Computer Vision and Pattern Recognition Conference*, pages 103–109, 1997.

[6] S. Mann and R. W. Picard. Virtual bellows: Constructing high quality stills from video. In *ICIP*, 1994.

[7] L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering system. In *Proc. of SIGGRAPH*, pages 39–46, 1995.

[8] S. Peleg and J. Herman. Panoramic mosaics by manifold projection. In *CVPR*, pages 338–343, 1997.

[9] H. S. Sawhney, S. Ayer, and M. Gorkani. Model-based 2D&3D dominant motion estimation for mosaicing and video representation. In *Proc. Intl. Conf. on Computer Vision*, pages 583–590, 1995. ftp://eagle.almaden.ibm.com/pub/cs/reports/vision/dominant_motion.ps.Z.

[10] H. S. Sawhney and R. Kumar. True multi-image alignment and its application to mosaicing and lens distortion. In *CVPR*, pages 450–456, 1997.

[11] C. C. Slama. *Manual of Photogrammetry*. Amer. Soc. of Photogrammetry, Falls Church, VA, 1980.

[12] R. Szeliski. Image mosaicing for tele-reality applications. In *IEEE Wkshp. on Applications of Computer Vision*, pages 44–53, 1994.

[13] R. Szeliski and H. Shum. Creating full view panoramic image mosaics and environment maps. In *Proc. of SIGGRAPH*, pages 251–258, 1997.

[14] L. A. Teodosio and W. Bender. Salient video stills: Content and context preserved. In *ACM Intl. Conf. on Multimedia*, 1993.

[15] VideoBrush. http://www.videobrush.com.

[16] Y. Xiong and K. Turkowski. Creating image-based VR using a self-calibrating fisheye lens. In *Proc. Computer Vision and Pattern Recognition Conference*, pages 237–243, 1997.

(a)

(b)

Fig. 9. (a) 184 frames captured using 3 successive horizontal 360° swipes are successfully mosaiced using local-to-global spherical parameters. The Mercator projection of the mosaic is cut into two pieces for clarity. (b) Same mosaic rendered onto a spherical surface.