High Resolution Video Mosaicing with Global Alignment

Roberto Marzotto, Andrea Fusiello, Vittorio Murino
Dipartimento di Informatica, Università degli Studi di Verona
37134 Verona, Italy
{andrea.fusiello,vittorio.murino}@univr.it

Abstract

Image composition (or mosaicing*) has attracted a growing attention in recent years as one of the main elements in video analysis and representation. In this paper we deal with the problem of global alignment and super-resolution. We also propose to evaluate the quality of the resulting mosaic by measuring the amount of blurring. Global registration is achieved by combining a graph-based technique – that exploits the topological structure of the sequence induced by the spatial overlap – with a bundle adjustment which uses only the homographies computed in the previous steps. Experimental comparison with other techniques shows the effectiveness of our approach.*

1. Introduction

Image mosaics have attracted a growing attention in recent years and they have been applied in many applications, like video stabilization and compression, background subtraction, virtual environments and panoramic photography.

Mosaics can be constructed by aligning and properly blending together partially overlapped images acquired by a camera. If it follows a bidimensional scanning pattern (like a zig-zag path), instead of the more common monodimensional pattern (as in a panning sequence), it is necessary to devise *global* registration methods which aim to exploit all available information. Global registration (or alignment) refers to the alignment of video frames taking into account (ideally) all the overlapping frames, and not just the consecutive ones. The motivation for this comes from the observation that even if the alignment of consecutive frames is accurate, simply concatenating local alignment models may lead to a gross global misalignment.

Many approaches for global registration have been proposed in the last years. In [2], the global consistency of the inter-frame alignment matrices is enforced by solving a linear system of equations. A weak point of this technique is that it can cope with an affine deformation only. In [1], the maximum likelihood estimate of the set of consistent homographies is computed given all the point matches. In [11], global registration is achieved by minimizing differences between ray directions going through corresponding points. The most closely related work are [10, 6], which both use a graph representation and cast the problem as the identification of the shortest path.

Our method for global registration consists of three principal steps: frame-to-frame registration, graph-based registration and simultaneous registration (i.e., bundle adjustment). First, only consecutive frames are registered. Then, a graph is built, whose vertices are the frames and edges link frame pairs for which a planar transformation is directly computed. The edge costs reflect the local registration accuracies. The graph representation allows us to search for the optimal path connecting every frame in the sequence to a chosen reference frame. The third and last step uses local constraints to determine the optimal global registration that minimizes an error function.

This work, albeit similar in some aspects to the above mentioned ones, differs substantially in many parts. In particular, a cost (error) function is proposed which leads to mosaics of better quality with fewer iterations as compared with those proposed by [11] and [6], according to the comparative analysis reported in Section 4. We also address the construction of super-resolution mosaics of high quality [4, 12]. Indeed, this results as a side effect of computing precise (sub-pixel) alignment between frames. This is possible because a sub-pixel motion defines a (non-uniform) sampling grid on the mosaic which is finer than the original pixels grid. Section 3 describes our approach to super-resolution. Another contribution of this paper is the proposal of an objective measure for evaluating the quality of mosaics. In Section 4.1 we describe this measure and give experimental support to this choice.

2. Mosaicing

Image mosaicing can be defined as the automatic alignment (or registration) of multiple images into larger aggregates [14].

Two pictures of the same scene are related by a (non-singular) linear transformation of the projective plane in two cases: i) the scene is planar or ii) the point of view does not change (pure rotation). In these cases, which can be summarized by saying that there must be no *parallax*, images can be composed together to form a *mosaic*.

Points are expressed in homogeneous coordinates, that is, 2-D points in the image plane are denoted as $\tilde{\mathbf{x}} = (x, y, 1)$ with $\mathbf{x} = (x, y)$ being the corresponding Cartesian coordinates.

A linear transformation of the projective plane, called a *homography*, is represented by a 3×3 matrix $\mathbf{H}$ such that $\tilde{\mathbf{x}}_i = \mathbf{H}_{ij}\tilde{\mathbf{x}}_j$, where $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$ are corresponding points in frame i and j respectively. Four points, provided that no three of them are collinear, determine a unique homography.

2.1. Frame-To-Frame alignment

Inter-frame homography computation is based on the Kanade-Lucas-Tomasi (KLT) tracker [7, 15], initialized with phase-correlation to reduce search range. Features are tracked through the video sequence and correspondences are used to compute homographies $\mathbf{H}_{i,i+1}$ between each pair of consecutive frames. As in [9], Least Median of Squares is used to be robust against tracking errors and features attached to moving objects. Given the set of *inlier* point matches, the homography computed with the least-squares method is affected by a statistical bias, because homogeneous coordinates are treated as if they were physical quantities without distinguishing measurement data from artificial constants. In order to obtain an optimal estimation of homography and reduce the instability of images mapping even with a small overlap between frames, we apply the method proposed by Kanatani [5], based on a renormalization technique.

These homographies are then concatenated to obtain a first mosaic. Any frame can be chosen as the reference one onto which register all the others.

The frame-to-frame registration consists in computing, for each frame i the homography $\mathbf{H}_{r,i} \triangleq \mathbf{H}^t_i$ that maps frame i onto the reference frame r, using recursively the homographies that links consecutive frame pairs:

$$\begin{cases} \mathbf{H}_k = I & \\ \mathbf{H}^t_i = \mathbf{H}^t_{i-1}\mathbf{H}_{i-1,i} & \text{if } i > r \\ \mathbf{H}^t_i = \mathbf{H}^t_{i+1}\mathbf{H}_{i+1,i} & \text{if } i < r \end{cases}$$

When the video sequence is long enough, this straightforward way to compose homographies yields an appreciable misalignment for the frames farther from the reference one. This is especially evident when the camera goes back onto a scene part previously seen. In this case, registration can take advantage from homographies linking non-consecutive frames and reduces the global misalignment error.

2.2. Graph-based alignment

This stage is a slight modification of [10].

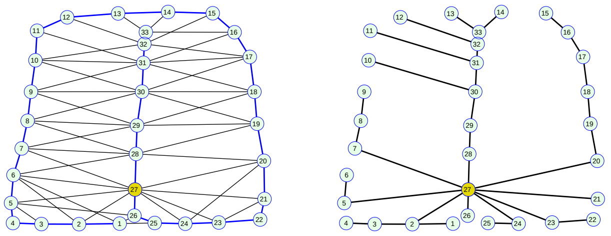


Figure 1. Left: graph for "S. Zeno" sequence. Nodes position is given by the centroid of the corresponding frame in the mosaic reference frame. The blue (bold) edges link consecutive frames in the sequence. Black (thin) edges are those added by our algorithm. Right: Spanning tree composed by all shortest paths from each node to the root.

The first step is to establish the frames which overlaps each other. The frame-to-frame alignment $\{\mathbf{H}^t_i\}$ gives a good approximation of the registration matrices, which allows to estimate the degree of overlap between each frame.

In principle one could compute the homography linking each overlapping pair, but as this is an expensive operation, it should be performed judiciously.

A graph is constructed, whose vertices are the frames and edges links overlapping frame pairs for which it makes sense to compute an homography directly. Initially, only consecutive frames in the sequence are connected. Then edges will be added incrementally, linking frames that (following [10]) i) have a significant overlap, and ii) creates a material shortcut between two vertices.

As an overlap measure we use the normalized distance between centroids:

$$\delta_{ij} = \frac{\max(0, |\mathbf{c}_i - \mathbf{c}_j| - |d_i - d_j|/2)}{\min(d_i, d_j)} \tag{1}$$

where $\mathbf{c}_i$, $\mathbf{c}_j$, d_i and d_j are the centroids and the diameter of the projection onto the mosaic of frames I_i and I_j, respectively. If $\delta_{ij} > 1$ there is no overlap. As for criterion ii), we compute

$$\gamma_{ij} = \frac{\delta_{ij}}{\Delta_{ij}} \tag{2}$$

where Δ_{ij} is the cost of the shortest path between node i and node j, in the graph with weights δ on the edges. γ_{ij}

varies in the range $[0, 1]$; the higher its value, the less influence it is likely to have pair (i, j) on the simultaneous registration. Please note that the value of γ_{ij} depends (via Δ_{ij}) on the graph's edges.

In order to choose which edges have to be added to the graph (i.e., which homographies have to be computed beside those linking consecutive frames) the following greedy algorithm is used:

```
$T := \{(i,j) | j = i+1\}$; % consecutive frames
$S := \{(i,j) | j > i+1\}$; % all pairs
for each $(i,j) \in S$ compute $\delta_{i,j}$ e $\gamma_{i,j}$;
$S := S \setminus \{(i,j) | \delta_{i,j} \geq s \vee \gamma_{i,j} \geq t\}$;
$L := T$ ;
while $S \neq \emptyset$;
    $e := \arg\min_{(i,j)\in S} \gamma_{i,j}$;
    $S := S \setminus \{e\}$;
    $L := L \cup \{e\}$;
    for each $(i,j) \in S$ compute $\gamma_{i,j}$
    $S := S \setminus \{(i,j) | \gamma_{i,j} \geq t\}$;
end while
```

For all the edges (i, j) in $L \setminus T$, the corresponding homography $\mathbf{H}_{ij}$ is computed directly from feature correspondences. As the two frames i and j are not consecutive, features can undergo severe perspective distortion. To overcome this problem and obtain a more accurate matching (KLT tracker is based on a translational model) the two frames are first transformed onto the reference frame (with $\mathbf{H}_i^t$ and $\mathbf{H}_j^t$ respectively), thereby compensating the distortion.

In the final graph (Fig. 1) we compute, for each frame i, the transformation $\mathbf{H}_i^s$ that aligns it with the reference frame by chaining homographies along the shortest weighted path from i to r, where edges are weighted with the mean squares residual of the homography computation.

This is equivalent to computing the (weighted) minimum spanning tree (MST) with the reference frame as root. The idea is that this yield a solution $\mathbf{H}_i^s$, which we call *MST solution*, that is less affected by errors accumulation than $\mathbf{H}_i^t$ because it is the product of (possibly) fewer low-residual factors.

2.3. Bundle adjustment

The bundle adjustment finds the solution $\{\mathbf{H}_i\}$ that minimizes the total misalignment of a pre-defined set of m grid-points on the mosaic. Let $\mathbf{x}_k$ be a grid-point and let L_k be the set of edges $(i, j) \in L$ such that $\mathbf{x}_k$ belongs to the overlap region between frame i and frame j. The error at the grid-point $\mathbf{x}_k$ is defined as:

$$E_k = \frac{1}{|L_k|} \sum_{(i,j)\in L_k} ||\mathbf{x}_k - \Pi(\mathbf{H}_i \mathbf{H}_{ij} \mathbf{H}_j^{-1} \tilde{\mathbf{x}}_k)||^2 \tag{3}$$

where Π transforms homogeneous coordinates into Cartesian (pixel) coordinates. Since we want to simultaneously minimize the error at all the grid points, we end up with a system of non-linear equations that can be cast as a least-squares problem:

$$\min_{\{\mathbf{H}_i\}} \sum_{k=1}^{m} E_k^2 \tag{4}$$

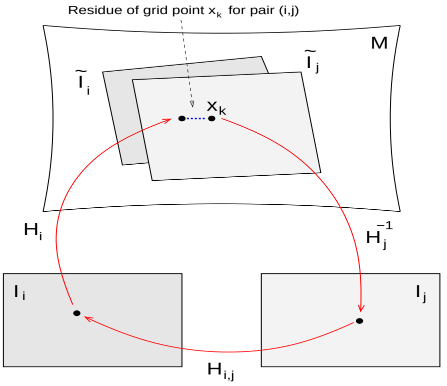


Figure 2. Simultaneous registration diagram.

The Levenberg and Marquardt algorithm[1] is used to solve Eq. (4), using $\{\mathbf{H}_i^s\}$ as the starting solution. Usually this is already a good solution and few iterations are needed to get to the global minimum. As suggested by [3], data standardization is carried out to improve the conditioning of the problem.

The rationale behind the design of our cost function is twofold:

- using a regular grid of points in the mosaic ensures a uniform distribution of the misalignment error *in the mosaic* and
- computing residuals in the mosaic reference frame is more correct than computing them in the images reference frame, because in this way we minimize a quantity which is more closely related to the perceived misalignment in the mosaic.

3. Super-resolution

Our approach to super-resolution was inspired by [12], where sub-pixel motion information of a global motion model is used to create mosaics with a resolution that is higher than the resolution of each single video frame. In [12] the mosaic is constructed with *source-scan* warping: each pixel of each frame is mapped onto the mosaic, and only if its position is close enough to an integer (± 0.2), its

[1] Available through the `lsqnonlin` MATLAB function.

color is assigned to the corresponding pixel in the mosaic. Holes are eventually filled by interpolation.

This approach is attractive as it is simple, and super-resolution arises as a by-product of mosaicing with sub-pixel accuracy. However, it has three drawbacks, which our method addresses: i) it contains an arbitrary threshold, ii) it may leave holes to be filled with interpolation (not a good idea when doing super-resolution) and iii) it suffers from the *folding* problem, i.e., a pixel in the mosaic may be assigned more than once.

Our super-resolution mosaic is built using *destination-scan* warping: for each pixel in the mosaic (which has a resolution greater than the single frames), find the corresponding position in each frame by backward mapping (with properly scaled transformations) and pick the color of the *nearest* pixel, over all the frames.

This “nearest pixel” strategy works well only if the registration is *very* accurate. In practice, a weighted strategy gives usually better results: backward-map the mosaic pixel in each frame, find the closest pixel and weigh its color with the inverse of the distance; finally, assign to the mosaic pixel the weighted average of the colors. The weighting function we used is $\log(\frac{1}{\sqrt{2}x})$, which is 0 when $x = 1/\sqrt{2}$ (the maximum distance to the closest pixel is bounded by $1/\sqrt{2}$) and goes to positive infinity as x approaches 0.

4. Results

4.1. Comparison of quality measure

The quality evaluation of mosaics is usually subjective, and it’s based on the perceived blurring. We propose to use an *objective* blurring measure, taken from the vast literature on focusing [13]. In particular, we chose the *energy of the image Laplacian*

$$\mathrm{EL}(I) = \frac{1}{n}\sum_{x,y}(\nabla^2 I(x,y))^2$$

as it is smooth and has a sharp maximum. We wanted to test the appropriateness of EL as a measure of the quality of the alignment, and to compare it with other focus operators, namely: Sum-Modulus-Difference (SMD), Sum-Modified-Laplacian (SML) and Tenengrad, as defined in [13, 8].

A synthetic sequence of 15 identical frames were generated by replicating an image composed by black and white bands. The inter-frame homographies (the identity) were perturbed by adding noise with increasing amplitude σ to their entries. 70 independent trials were performed for each noise level. In each trial a mosaic was constructed and the focus operators were applied to it. The output of each operator was averaged over the 70 trials. The result is shown in Fig. 3.

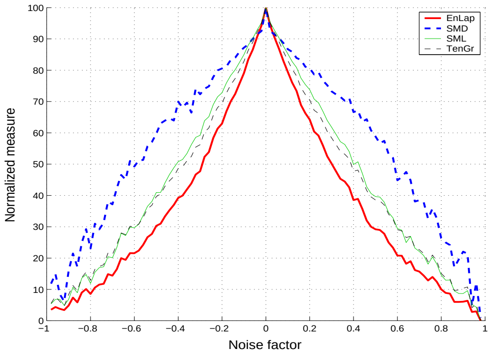


Figure 3. Output of several focus operators vs homography perturbation.

All the different operators have the maximum for $\sigma = 0$, but the EL is smoother (i.e. has no local maxima) and sharper.

4.2. Comparison with other cost functions

The bundle adjustment described in Sec. 2.3 is based on an original cost function that we devised. Other cost functions had been introduced in [11], and [10]. In this section we briefly recall the description of this two cost functions and in the next section we compare results. It must be clear that we are not comparing our mosaicing solution with those presented in [10] and [11], which are complex and differs from ours in many aspects other than the cost function.

The cost function used in [10] has the following form[2]:

$$\min_{\{\mathbf{H}_i\}} \sum_{ij\in L} E_{ij}^2 \tag{5}$$

where

$$E_{ij} = \sum_{k=1}^{4} \|\mathit{\Pi}(\mathbf{H}_i\tilde{\mathbf{u}}_{kij}) - \mathit{\Pi}(\mathbf{H}_j\mathbf{H}_{ij}^{-1}\tilde{\mathbf{u}}_{kij})\|^2 \tag{6}$$

The error E_{ij} penalizes the inconsistencies between frame-to-mosaic homographies ($\mathbf{H}_i$ e $\mathbf{H}_j$) and the inter-frame homography ($\mathbf{H}_{ij}$). Points $\mathbf{u}_{kij}$ (with $k \in \{1..4\}$) are the four vertices of the overlap region between I_i and I_j represented in the reference frame of I_i.

In [11] the bundle adjustment consists in solving the following problem:

$$\min_{\{\mathbf{H}_i\}} \sum_{ij\in L} \sum_{k\in N_{ij}} E_{kij}^2 \tag{7}$$

[2] In the original formulation a regularization term was added in order to fix the global reference frame. In our case the reference frame is fixed arbitrarily.

where

$$E_{kij} = \|\mathbf{m}_{ki} - \Pi(\mathbf{H}_i^{-1}\mathbf{H}_j\tilde{\mathbf{m}}_{kj})\| \qquad (8)$$

The set $N_{i,j}$ contains the indices of the "good" correspondences $(\mathbf{m}_{ki}, \mathbf{m}_{kj})$ between images I_i and I_j. In our implementation these are the inlier correspondences produced by the robust homography estimation. E_{kij} is the residual for feature k between images I_i and I_j computed in the reference frame of image I_i.

The main difference between these two approaches is that in the former, as in ours, the information contained in point correspondences gets distiled into homographies, whereas in the latter all the original corresponding points are used in the bundle adjustment.

4.3. Experiments

All the video sequences were taken with a digital handheld camera. Frames are 338 × 280 pixel. Radial distortion have been preliminary compensated by calibration [16]. All the sequences and more results are available on the web[3]. Experiments have been carried out on a Pentium 4 1500 MHz; the code is written in MATLAB and C.

In the first example we deal with a panoramic mosaic. The full facade of S. Zeno cathedral was imaged while rotating the camera (approximately) around its optical center.

The final mosaic is shown in Fig. 4. As a blending operator for standard resolution mosaics we used the average. Feathering could be used to overcome border effect [11] and median could be used to get rid of (fast) moving objects [9].

The improvement of the global registration is particularly evident in the area at the bottom center of the mosaic (Fig. 5).

The second example is a planar scene, namely a map of Europe, that was imaged by translating the camera. Figure 6 depicts the resulting mosaic whereas Fig. 7 shows a detail where the benefit brought by the bundle adjustment can be appraised.

Figure 8 shows the distribution of alignment error over grid points. After bundle adjustment the error is smaller and more uniformly distributed.

Figure 9 shows quality measures for mosaics obtained from frame-to-frame solution, MST solution and bundle adjustment, starting from both frame-to-frame and MST and with three different cost functions.

It can be noted that the graph-based registration always improves over frame-to-frame registration and that the quality of the final mosaic is always better when starting from the MST solution as opposed to starting from frame-to-frame solution.

Our cost function gives better results than (6) and perform only slightly worse than (8), but our approach takes

[3] http://profs.sci.univr.it/˜fusiello/demo/hrm

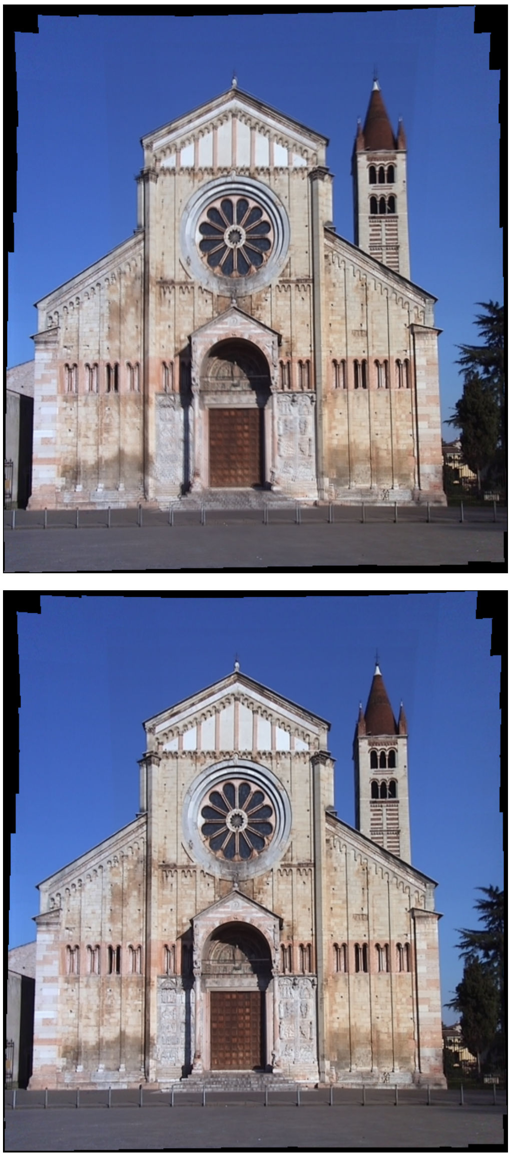


Figure 4. Mosaics of "S. Zeno" before (top) and after (bottom) global alignment.

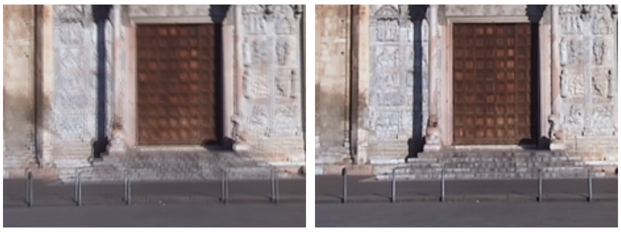


Figure 5. Detail of the mosaics before (left) and after (right) global alignment.

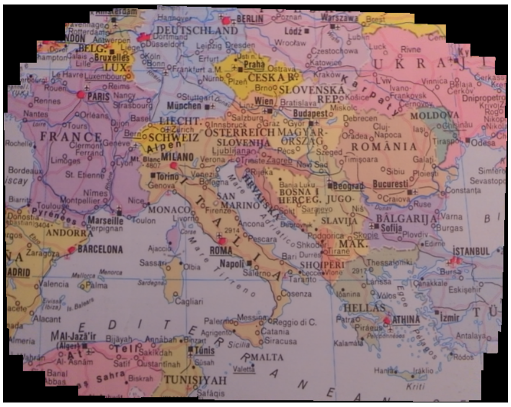


Figure 6. Mosaic of the map of Europe after global alignment.

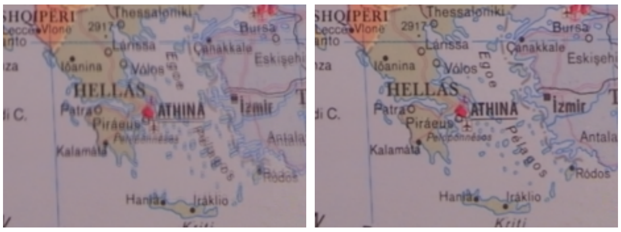


Figure 7. Detail of the mosaic before (left) and after (right) global alignment.

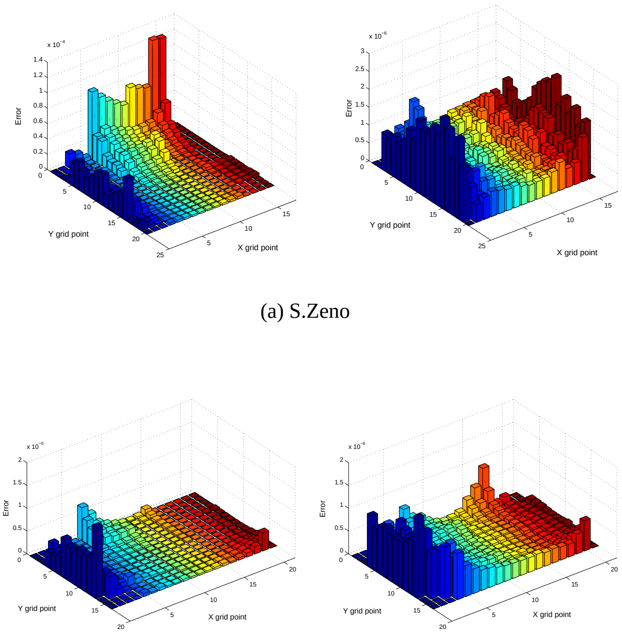
fewer iterations to converge (see Table 1). Indeed, it stops after few iterations, while the others reach the maximum number of iterations (set to 30).

Method	Iter	Time (s)	Method	Iter	Time (s)
Our	8	116	Our	6	90
Sawhney	30	168	Sawhney	30	134
Szeliski	30	257	Szeliski	30	194

Table 1. Computation times for "S. Zeno" (left) and "Europe" (right)

Figure 10 demonstrates that EL consistently increases with the number of iterations, even if it is not directly related to the error measure that is minimized. In this case we let our algorithm to run for 30 iterations, behind convergence.

Results concerning super-resolution are visible on the web for subjective evaluation.



(a) S.Zeno

(b) Europe

Figure 8. Alignment errors on mosaic grid-points before (left) and after (right) bundle adjustment (please note that scale on the vertical axes are different).

5. Conclusions

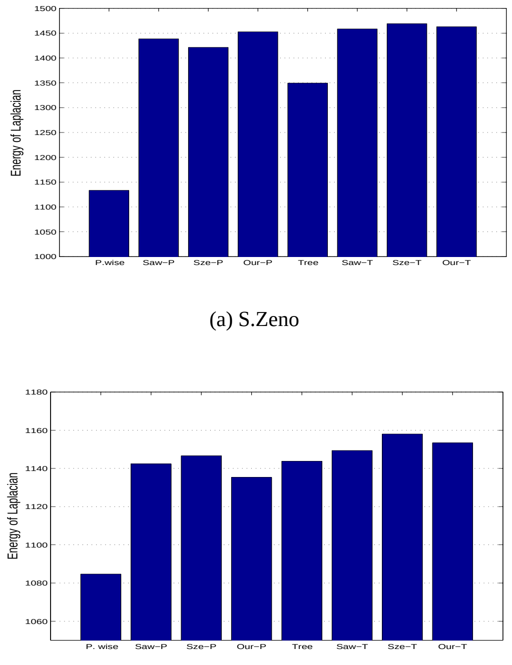
In this work we presented a new technique for automatic construction of panoramic and planar mosaics from video sequences using projective transformations. Novel contributions include global registration and super-resolution. A new measure for mosaic quality estimation was also introduced. Experiments confirm the goodness of our approach, which yields mosaics of better quality with fewer iterations compared to existing methods.

Acknowledgments

A. Colombari provided the implementation of the KLT tracker and contributed with useful discussions. The implementation of the algorithm described in [5] is available on the web by courtesy of Naoya Ohta and Shimizu Yoshiyuki.

References

[1] D. Capel and A. Zisserman. Automated mosaicing with super-resolution zoom. In *Proceedings of the IEEE Con-*



(a) S.Zeno

(b) Europe

Figure 9. Value of EL for different solutions. From left to right: frame-to-frame, bundle adjustment starting from frame-to-frame using cost function (6)**,** (8)**, and** (3)**, MST, bundle adjustment starting from MST using cost function** (6)**,** (8)**, and** (3)**.**

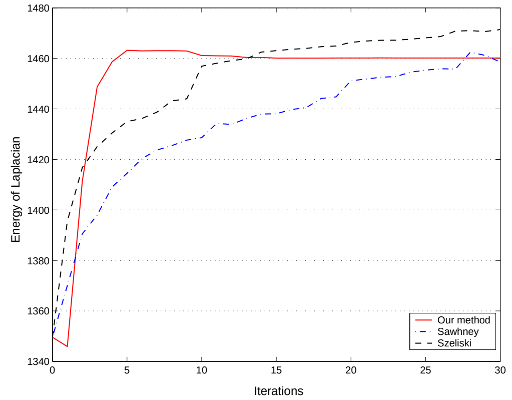


Figure 10. Value of EL vs iteration number in the bundle adjustment of "S. Zeno" starting from MST using three cost functions.

ference on Computer Vision and Pattern Recognition, pages 885–891, 1998.

[2] J. Davis. Mosaics of scenes with moving objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 354–360, 1998.

[3] R. I. Hartley. In defence of the 8-point algorithm. In *Proceedings of the IEEE International Conference on Computer Vision*, 1995.

[4] M. Irani and S.Peleg. Improving resolution by image registration. *Graphical Models and Image Processing*, 53(3):231–239, May 1991.

[5] K. Kanatani and N. Ohta. Accuracy bounds and optimal computation of homography for image mosaicing applications. In *International Conference on Computer Vision*, volume 1, pages 73–79, Sept. 1999.

[6] E. Kang, I. Cohen, and G. Medioni. A graph-based global registration for 2D mosaics. In *Proceedings of the International Conference on Pattern Recognition*, pages 257–260, Barcellona, Spain, September 2000.

[7] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1981.

[8] N. Nathaniel, N. Aun, and A. Marcelo. Practical issues in pixel-based auto-focusing for machine vision. In *Proceedings of the 2001 IEEE International Conference on Robotics and Automation*, pages 2791–2796, May 2001.

[9] F. Odone, A. Fusiello, and E. Trucco. Layered representation of a video shot with mosaicing. *Pattern Analysis and Applications*, 5(3):296–305, August 2002.

[10] H. S. Sawhney, S. Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. In *Proceedings of the European Conference on Computer Vision*, volume 1407, pages 103–119, 1998.

[11] H.-Y. Shum and R. Szeliski. Construction of panoramic image mosaics with global and local alignment. *International Journal of Computer Vision*, 36(2):101–130, 2000.

[12] A. Smolic and T. Wiegand. High-resolution video mosaicing. In *Proceedings of the IEEE International Conference on Image Processing*, Thessaloniki, Greece, 2001.

[13] M. Subbarao, T. Chio, and A. Nikzad. Focusing techniques. *Optical Engineering*, pages 2824–2836, 1993.

[14] R. Szeliski. Video mosaics for virtual environments. *IEEE Computer Graphics and Applications*, 16(2):22–30, March 1996.

[15] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, Pittsburg, PA, April 1991.

[16] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.