Panel Introduction
Video Registration: Key Challenges and the Potential Impact of Their Solutions to the Field of Computer Vision

Mubarak Shah
Dept. of Computer Science
University of Central Florida
Orlando, FL 32816
shah@cs.ucf.edu

In order to review past accomplishments and discuss future challenges, we organized a panel session during the workshop. Four panelists were invited to participate: Steve Blask from Harris Corporation, Lisa Brown from IBM, Harpreet Sawhney from Sarnoff Corporation, and Rick Szeliski from Microsoft. The panelists were asked to select three or four questions from the following ten questions and express their views. In this section, I will discuss each question, and express some of my opinions. The next three sections deal with the opinions of the three panelists.

1 *What has been accomplished in video frame-to-frame registration in the context of mosaics, panoramas, etc., and what are the new challenges?*

One success story of motion analysis research is the estimation of global frame-to-frame motion. Traditional efforts to estimate pixel-wise optical flow have met with mixed success. However, for the estimation of global motion, the information in all pixels is used to estimate the global transformation, like affine, projective, or psuedo-perspective. Such transformations can then be used to align video frames to generate mosaics or panoramas. This has been a very active area during the last few years, and continues to be a hot area in the context of computer graphics, visualization, surveillance, segmentation, etc.

2 *What is the role of video registration in object-based segmentation of images captured by a moving camera, and how far are we from automatic segmentation of video objects for an arbitrary scene in the context of MPEG-4?*

Object-based segmentation of video is very important for video compression, video understanding, etc. In particular, segmentation of video containing both object and camera motion is pretty complex. One obvious solution is to first estimate the camera motion and compensate for the motion to generate video with no camera motion. This can then be used to segment moving and stationary objects. However, estimation of camera motion with large local motion is a pretty difficult problem. The other alternative is to segment and track each individual object throughout the video, without necessarily estimating a global motion.

3 *Does Engineering Support Data (telemetry) like the DEM and the camera orientation and location really help in registration of video frames with the reference image? What are the hard problems in this area?*

The traditional structure from motion problem in computer vision deals with the recovery of camera translation and rotation and the scene depth using two or more images. Even though lots of theoretical work has been done in this area during the last two decades, it is still not possible to robustly solve the structure from motion problem for any general scene, since there are too many unknowns and the problem is non-linear in nature. If the 3-D rotation and translation between the video and the reference image can be recovered, the registration of the video image with the reference image becomes trivial. When nothing else besides the two images is known, the registration is the most complex. However, when additional information is available, the registration can be simplified and more accurate results can be obtained. In some cases, additional information (metadata or telemetry) about how images were taken, like the location and orientation of the camera and the Digital Elevation Map (DEM), is readily available and can be used as an aid in the registration process. For pictures taken by various satellites, like LANDSAT, SPOT, or IRS, telemetry data is widely available. In the context of the DARPA Airborne Video Surveillance (AVS) program, each video frame contained telemetry data, including aircraft longitude, latitude, heading, and velocity.

4 *Is registration of 3-D data (e.g. CT with MR) easier than registration of video (2-D) data with 3-D data (e.g. overlay of video on CT)? Or vice versa?*

Video imagery projects a 3-D scene on a 2-D image plane, therefore one dimension is lost during the perspective projection. Moreover, 2-D video imagery does not contain any 3-D information. On the other hand, both CT and MR are in 3-D nature. Therefore, registration of video with 3-D data requires estimation of the 3-D pose of the video camera in order to overlay video on CT or MR.

5 Is the wide baseline stereo problem solved now?

Several good algorithms exist for small baseline stereo, such as graph cuts, layers, and SAD. However, some very interesting applications need to obtain disparity maps over a wide baseline, such as view morphing and reconstruction from multiple images. The most popular approaches use the affine model, which works very well when the image warping includes scaling, shearing and small rotation, since all of them can be approximated as linear expansion of the first order gradients. However, the affine model fails to track if a large rotation between the two frames is introduced. Since the rotation component is non-linear indeed, it is very difficult to capture the non-linear part by employing the linear expansion of horizontal and vertical image gradients I_x and I_y.

6 *Does correlation still play an important role in registration? Has anything new happened in correlation during the last 50 years?*

Correlation is probably the oldest method for registering two images, and it remains one of the most popular methods used in industry, due to its simplicity and robustness. One big drawback of correlation is its large computational complexity, therefore several efficient hardware and software approaches have been proposed to speed up the computation. Currently, new approaches that compute the correlation of image histograms (distributions) instead of individual pixel intensities are also being explored. Correlation of raw intensity or color values may be more sensitive to image noise, therefore correlation of the gradient or texture features offers more attractive alternatives.

7 *Is video registration harder than visual tracking? Why or why not?*

Tracking using a static camera does not require image registration. Tracking using a moving camera is much more complex, since the global motion caused by the camera motion must be differentiated from the local object motions. As mentioned earlier, one possible approach is to first estimate and compensate for the global camera motion, and then perform tracking in stabilized video. However, tracking in videos acquired by moving cameras can also be performed without global motion compensation. A simple way to achieve this is to perform object detection (object segmentation) in every frame, and then solve the motion correspondence between detected object regions. One possible class of methods for object detection is those based on active contours.

8 What is the role of image features in video registration?

Registration can be performed by using raw intensity or color values directly, therefore these methods have been called *direct* methods. Another method first detects features in each frame, then correspondence between these features can be solved to register images. The most common features include edges, lines, corners, interest points, line intersections, roads and building structures. The difficulty in the feature-based approaches to registration is the robust detection of features in images of featureless scenes like deserts and forests.

9 What are the next most important problems which need to be solved in video registration?

Registration of video in the presence of large local motion and parallax is still hard problem. Registration of video with site models is another interesting problem. Finally, spatial and temporal registration of non-overlapping video sequences is pretty challenging, and will find many uses in the future.

10 *What have been the most successful solutions so far and the most successful approaches?*

The use of all image information in a least squares fashion to estimate frame-to-frame motion is probably one of the most successful approaches in computer vision, compared to estimation of pixelwise optical flow. As a result, mosaics and panoramas generated from video sequences using estimated global motion are very impressive.

ImageVideo Registration Retrospective

Lisa Brown
IBM T.J. Watson Research Center
New York
lisabr@us.ibm.com

1. Introduction

A decade ago, as part of my candidacy requirements, I wrote a survey of image registration methods [4]. Since that time, there has been widespread application of image registration methods and an extensive body of new research. Image registration continues to be a critical part of almost all computer vision applications: tracking, modeling, pose estimation, shape estimation, etc. Given the enormous increase in computational speed and ten years of research, what have we learned about this problem? What are the most successful approaches to image and video registration problems?

In 1992, it was possible to categorize most image registration problems into four major groups based on whether the images to be registered came from different sensors, different viewpoints, at different times, or from a combination of a reference image or model and a newly acquired image. We refer to these four types of registration problems as: multimodal, viewpoint, temporal, and template registration, respectively. In addition, registration problems can also be usefully distinguished based on the dimensionality of the data, for example, the registration of 2D to 2D, 2D to 3D, 3D model to video, or video to video. Today, we find that researchers are able to tackle more difficult registration problems often involving a combination of the four fundamental groups and at the same time dealing with higher dimensional data.

Video registration problems generally fall into one of three categories:

1 Registration of video to reference imagery or 3D models. Examples include geo-registration, i.e., registration of aerial video to reference imagery with high geodetic accuracy [2], registration of physical patient space or pre-operative 3D medical data with operative video data (such as laparoscopy, laryngocoscopy, fluoroscopy, or ultrasound) to assist in minimally invasive surgery or other image-guided procedures [12], [26]. Other examples include urban model building or mosaicking (see Category (3) below). Another interesting instance is face sequence matching. In this application face recognition is performed on a video sequence [23]. These problems are all typically a form of template registration with high data dimensionality (templates are often 3D models, input is video). The high dimensionality inevitably adds temporal and viewpoint issues as well.

2 Video to video registration, i.e., finding a video clip in a longer sequence. Examples include video copy detection [7], [1, 20], or video content retrieval [30] and synchronizing multiple video cameras [5, 15]. All of these problems involve temporal registration, the former involves subtle sensor differences (i.e., different models/brands of video cameras) and the latter includes viewpoint registration. Another example is multi-modal fusion, i.e., two video sensors of different modalities. However, the majority of current research in this area is still image-based or image/video based.

3 Frame-to-frame registration such as camera motion estimation and video enhancement. This category includes all manner of video quality improvements and virtual visualizations. Quality improvements include exposure compensation, lens distortion correction

[22, 9] and video stabilization [8, 14]. Virtual visualizations include the ability to superimpose computer-generated imagery on dynamic scenes [6], the creation of mosaics [10, 18, 8, 22] and 3D model visualizations [25, 11]. Frame-to-frame registration primarily involves viewpoint registration. Quality improvements add additional sensor models (multimodal registration) while visualizations often require registration to a reference (i.e. template registration) and other assumptions about the 3D world. Some of the methods that create mosaics, panoramas and 3D models, overlap with the first category since they perform frame-to-frame registration but they do this concomitantly with creating a reference frame/model.

2. Image/Video Registration Framework

In order to organize the extensive range of image registration methods, it is useful to establish the relationship between the variations in the images and the type of registration technique, which can most appropriately be applied. Three major types of variations are distinguished. The first type is the variations due to the differences in the acquisition, which cause the images to be misaligned. To 'register' images, typically a spatial transformation is found which will remove these variations. The class of transformations, which must be searched to find the optimal transformation, is determined by knowledge about the variations of this type. We call these variations, Type I. The transformation class determines the search space, which in turn, influences the general technique that should be taken.

The second type of variations, Type II, is also due to differences in acquisition, but are difficult to model, such as differences due to variations in lighting, atmospheric conditions or differences in sensor responses. This type usually affects intensity values, but they may also be spatial, such as differences in perspective distortions due to different viewpoints. Type I variations are distortions which are modeled, while Type II variations are those which it was not possible to model.

The third type of variations, Type III, arises from differences in the images that are of interest such as object movements, growths (deformations) or other scene changes. Variations of the second and third type are not directly removed by registration, but they make registration more difficult since an exact match is no longer possible. In particular, it is critical in some applications that variations of the third type are not removed.

Knowledge about the characteristics of each type of variation affects the choice of (1) search space and search strategy, (2) feature space, and (3) similarity measure that will make up the final technique. All registration techniques can be viewed as different combinations of these three choices. This framework is useful for understanding the merits, limitations and relationships between the wide variety of existing techniques and for assisting in the selection of the most suitable technique for a specific problem. Using this framework, it is possible to review the extensive body of literature in image and video registration and summarize the capabilities and trends of the state-of-the-art in registration.

For specific fields, such as medical imaging, a specialized framework similar to this, but more extensive, has been developed in order to organize the existing methods for the practitioner. As put forth by [17], the framework for medical image registration, includes several categories that help differentiate the various problem types. These include the data dimensionality, the modality, the anatomical part, and whether the registration is between subjects, between two images of the same subject or between an atlas and a subject. Also, specific to medical image registration, methods are distinguished based on whether they are extrinsic, intrinsic or non-image based and by the level of user interaction required. Extrinsic methods rely on artificial objects, such as markers, that are attached to the patient. Historically, medical image registration has often relied on extrinsic methods but these are typically limited to cases where rigid transformations are sufficient and provisions can be made in the pre-acquisition stage. Non-image based methods refer to methods in which the coordinate spaces of the sensors are pre-calibrated to each other. Intrinsic methods have become increasingly more prevalent. An interesting aspect of intrinsic techniques is the natural breakdown of methods into three categories: those that rely on landmarks or features, surfaces, or raw intensity information.

By organizing the literature around this taxonomy it is possible to determine the best method that is suitable for a specific application. An even more specialized taxonomy was developed for computer-aided surgery [12] and for digital subtraction in dental radiography [16]. Such taxonomies enable someone to elegantly organize the extensive body of work, which has been developed. It is useful not only for the user, but also for the developer. In reviewing the literature, it is possible to narrow the search directly to the relevant works, rather than being limited by searching for key words and from a few related papers. Similarly, it is easier to analyze the relationship between other methods.

Based on a survey of several recent papers and a categorization using this framework of registration problems, it is possible to recognize several trends within the field of image and video registration. These trends will be briefly described for each of the 3 major aspects of registration problems: search space (transform class) and search strategy, feature space, and similarity measures. At the same time, several of the questions posed at the Video Registration Workshop '01 will be addressed.

3. Search Space and Strategy

As mentioned above, during the last ten years, researchers have started to address problems with higher dimensional data, both 3D and video. These problems combine the issues regarding images taken at different times, from different sensors, from different viewpoints, and which are aligned to reference models or atlas imagery. In medicine, there has been a definite shift from extrinsic to intrinsic methods, from more user interaction to less. In general, the class of transformations is more complex; many sensor distortions and differences are currently modeled and a wide body of literature now exists on fully non-rigid or elastic transformations particularly for medical imaging [21, 27]. In video to video registration, it is necessary to align image data both spatially and temporally. The addition of temporal cues has been found to be a powerful cue in aligning video [5].

In other words, researchers are addressing problems in which more and more of the variations between images are modeled. Several of the distortions previously categorized as Type II, are now categorized as Type I. For example, [22] models lens distortion as well as 2D/3D view transformations in the creation of mosaic images. In my own work in multi-modal medical image registration, specific sensor relationships were modeled (such as radiographic film characteristics, the process of x-ray projection, and partial volume effects in computed tomography) in order to optimally compare the pixel values from different medical sensors [3]. Blask et al. register aerial imagery to a 3D elevation model in a two stage process: initially correspondence between images is computed using local correlation patches, but the final registration is based on the correction of an elaborate sensor model whose parameters are physically measured [2]. [6] has explored the registration of fully dynamic scenes (such as crowds of people or scenes of water) by modeling the time series variation of individual pixels. Registration methods have become increasingly more accurate as they model more of the distortions between images and more precisely capture the relationships between images.

One of the key ways to efficiently find complex transforms with a large number of parameters is to implement a progressive complexity strategy. This strategy entails dividing the optimization process into a sequence of steps with increasing complexity. For example, in the creation of an image mosaic [22], initially a 2D translation is computed for each image. Using this transform to re-align the images, a 2D affine transform is then found for each image. Finally, the affine parameters are used as an initial estimate in order to compute a projective transform with an added global lens distortion correction transform for each image. This technique improves the speed and convergence to local minima by providing a good initial estimate for each stage.

4. Feature Space

Since more and more of the variations between images are modeled, high-level feature extraction has become less relevant. This has been clearly evident in the algorithms developed more recently. More than ever, registration is implemented using raw pixel intensity. When it is possible to use pixel-based information, registration methods are more powerful. According to [17], in medical imaging, methods that use full image content are gradually setting the standard for registration accuracy.

There are two major exceptions to this phenomenon. First of all, when speed is an issue, feature based methods are often necessary. In medical imaging applications such as radiotherapy treatment or intra-operative procedures, feature-based methods are more frequently encountered. These tend to extract surface information in order to efficiently align meaningful anatomic structures. In a survey of head tracking methods (for human computer interaction) [28] examines the trade-off between accuracy and speed for feature vs. template-based tracking methods. Although head tracking also involves segmentation, it is basically a frame-to-frame registration problem. Indeed, most tracking applications are a special case of registration.

The other rationale for extracting features occurs when it is not possible to model all the distortions between images, i.e., distortions of Type II or III. Features can be used to accurately localize positions that are more likely to be accurate. A common example occurs in wide baseline stereo. Registration techniques are used to fit a small number of parameters but the three dimensional structure of the objects in the scene is not known. An excellent way to handle this problem is to find the set of corresponding features. Finding such descriptors has been systematically developed over the past five years - descriptors that are invariant to affine and photometric transformations in order to achieve viewpoint invariance [24].

5. Similarity Measures

Similarity measures are used to evaluate the similarity between images. These measures depend on the choice of feature space. For raw intensity, the measure might be the sum of the absolute intensity differences or most commonly, the cross-correlation. For images in which color information is significant, histogram intersections are sometimes computed. For landmark/feature point, edge or surface based techniques the measure typically involves the minimal Euclidean distance between the points/edges/surfaces. Notice that for these types of features, similarity measures compute the spatial distance while for raw intensity or other pixel/voxel information, the measure deals with the scalar data such as intensity, color, or other physical properties.

As registration methods have become more sophisticated, transformation classes have become more complex. This has eliminated a great deal of uncorrected distortions (Type II) and decreased the need for feature extraction. The resulting voxel-property based registration methods rely on cross-correlation or related similarity measures. Cross-correlation has become the standard procedure because it can be efficiently computed and is the optimal measure if the remaining distortions are white noise. On the other hand, cross-correlation has its limitations. Global maximization is a difficult problem since many local minima exist. This is exacerbated by interpolation, which is necessary to find transformations at sub-voxel precision.

For monomodal registration, the correlation coefficient is frequently used, often a Laplacian pyramid to improve matching efficiency and insensitivity to lighting variation. For multimodal problems other measures are needed. A particularly well-suited measure for the general multi-modal registration problem is based on statistical correlation or mutual information (sometimes called relative entropy)[29, 19]. With this measure, the statistical relationship is maximized, i.e., the particular correspondence of image intensities between the two modalities is not presupposed but the proper transformation will maximize the degree of dependence between image intensities. This is useful if there is no a priori information regarding the correspondence. On the other hand, the relationship between sensor intensities is not always global and may depend on resolution [13].

6. Conclusions

The need to register images and videos is as prevalent today as ever before. Applications in surveillance, copy detection, human computer interaction, content retrieval, inspection, enhanced visualizations, and medical imaging are only a few of the many examples in which registration plays an important role. Over the last 10 years, the most obvious improvement in registration methods is their ability to deal with higher dimensional data including 3D to 3D, video to video, and 3D to video. In addition, methods today model more of the distortions between images, including sensor distortions, nonrigid deformations, and large viewpoint variations. The use of more sophisticated models has decreased the need for complex feature extraction and methods that rely on spatial similarity measures. The most successful methods, i.e. the most accurate, use voxel properties and often solve for transformations with many parameters. This is often efficiently optimized using a progressively complex transformation strategy.

Registration methods can be conveniently categorized using a simple framework which describes the problem type, search space and strategy, feature space and similarity measure. For specific applications, it is useful to organize methods by additional characteristics. These frameworks can be very helpful for comparing and systematically evaluating methods, selecting methods that are most suitable for specific problems, and sharing methodologies across domains.

References

[1] D. A. Adjeroh, M. C. Lee, I. King, "A Distance Measure for Video Sequences," Computer Vision and Image Understanding, Vol. 75, Nos. 1&2, p25-45, July/August 1999.

[2] Richard W. Cannata, Mubarak Shah, Steven G. Blask, John A. Van Workum, "Autonomous Video Registration Using Sensor Model Parameter Adjustments", Applied Imagery Pattern Recognition Workshop (AIPR) 2000, Cosmos Club, Washington D.C., Oct 16-18, 2000.

[3] L. Gottesfeld Brown, "Registration of Multi-Modal Medical Images: Exploiting Sensor Relationships," Dissertation, Computer Science Dept., Columbia University, 1996.

[4] L. Gottesfeld Brown, "A Survey of Image Registration Techniques," ACM Computing Surveys, Vol. 24, No. 4, p325-376, December 1992.

[5] Y. Caspi and M. Irani, "A Step Toward Sequence-to-Sequence Alignment," IEEE Conf. on Computer Vision and Pattern Recognition," Vol. 1, Hilton Head Island, S.C., p682-689, June 13-15 2000.

[6] A.W. Fitzgibbon, "Stochastic Rigidity: Image Registration for Nowhere-static Scenes," IEEE Int'l Conf. Computer Vision, Vancouver, BC, p662-669, July 9-12, 2001.

[7] A. Hampapur and R. Bolle, "Comparison of Distance Measures fro Video Copy Detection," Proc. of the Int'l Conf. on Multimedia and Expo, Japan, August 2001.

[8] M. Hansen, P. Anandan, K. Dana, G. van der Wal, P. Burt, "Realtime Scene Stabilization and Mosaic Construction," Proc. Image Understanding Workshop, Vol. 1, Monterey, CA, p457-65, 1994.

[9] J. Helferty, C. Zhang, G. McLennan, W. Higgins, "Videoendoscopic Distortion Correction and Its Application to Virtual Guidance of

Endoscopy," IEEE Trans. on Medical Imaging, Vol. 20, No. 7, p605-617, July 7, 2001.

[10] Y.S. Heung, R. Szeliski, "Systems and Experiment Paper: Construction of panoramic Image Mosaics with Global and Local Alignment," Int'l Journal Computer Vision (Netherlands) Vol. 36, No. 2, p101-30, 2000.

[11] S. Hsu, S. Samarasekera, R. Kumar, H. Sawhney, "Pose Estimation, Model Refinement, and Enhanced Visualization Using Video," IEEE Conf. on Computer Vision and Pattern Recognition," Vol. 1, Hilton Head Island, S.C., p488-495, June 13-15 2000.

[12] G. Ionescu, S. Lavallee, J. Demongeot, J., "Automated Registration of Ultrasound and CT Images: Application to Computer Assisted Prostate Radiotherapy and Orthopedics," Medical Image Computing and Computer Assisted Intervention - MICCAI '99, Vol. 1679, p768-77, Cambridge UK, Sept 19-22, 1999.

[13] M. Irani and P. Anandan, "Robust Multi Sensor Image Alignment," Sixth International Conference on Computer Vision, Bombay, India, p959-66, January 1998.

[14] J.S. Jin, Z. Z. Zhu, G. Xu, "Digital Video Sequence Stabilization Based on 2.5D Motion Estimation and Inertial Motion Filtering," Real Time Imaging (UK) Vol. 7, No. 4, p357-65, August 2001.

[15] L. Lee, R. Romano, G. Stein, "Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame," IEEE Trans. on Pattern Analysis and Machin Intelligence, Vol. 22, No. 8, p758-767, August 2000.

[16] T.M. Lehmann, H-G Grndahl, and D.K. Benn, "Computer-based Registration for Digital Subtraction in Dental Radiology," Dentomaxillofacial Radiology, Vol, 29, p323-346, 2000.

[17] J.B. Antoine Maintz and Max. A. Viergever, "A Survey of Medical Image Registration," Medical Image Analysis, Vol, 2, No. 1, p1-36, 1998.

[18] A. Mittal, D. Huttenlocher, "Scene Modeling for Wide Area Surveillance and Image Synthesis," IEEE Proc. Computer Vision and Pattern Recognition, Hilton Head Island, SC, Vol. 2, p160-167, June 13-15, 2000.

[19] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, " Multimodality Image Registration by Maximization of

Mutual Information," IEEE Transactions on Medical Imaging, Vol. 16, No. 2, p187-198, April 1997.

[20] R. Mohan, "Video Sequence Matching," Proc. of the 1998 IEEE Conf. on Acoustics, Speech and Signal Processing, Vol. 6, p3697-700, Seattle WA, May 12-15, 1998.

[21] M. Otte, "Elastic Registration of fMRI Data Using Bzier-Spline Transformations," IEEE Trans. on Medical Imaging, Vol. 20, No. 2, p193-206, February 2001.

[22] H.S. Sawhney and R. Kumar, "True Multi-Image Alignment and Its Applications to Mosaicing and Lens Distortion Correction," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.21, No. 3, p235-243, March 1999.

[23] S. Satoh, "Comparative Evaluation of Face Sequence Matching for Content-based Video Access," Proc. 4th Int'l Conf. on Automatic Face and Gesture Recognition" Grenoble, France, p163-8, March 28-30, 2000.

[24] F. Schaffalitzky and A. Zisserman, "Viewpoint Invariant Texture Matching and Wide Baseline Stereo," Conf. on Computer Vision, Vancouver, BC, Vol. II, p636-643, July 2001.

[25] I. Stamos and P.K. Allen, "Automatic Registration of 2-D with 3-D Imagery in Urban Environments," Int'l Conf. on Computer Vision, Vancouver, BC, Vol. II, p731-736, July 2001.

[26] J.D. Stefansic, et. al., "Registration of Physical Space to Laparoscopic Image Space for Use in Minimally Invasive Hepatic Surgery," IEEE Trans. on Medical Imaging, Vol. 19, No. 10, p1012-1023, October 2000.

[27] J.-P. Thirion, "Non-rigid Matching using Demons," Proc. Computer Vision and Pattern Recognition, San Francisco, CA, p245-251, June 18-20, 1996.

[28] K. Toyama, "Prolegomena for Robust Face Tracking," Workshop on Automatic Facial Image Analysis and Recognition Technology (ECCV 98).

[29] P. Viola and W.M. Wells III, "Alignment by Maximization of Mutual Information," Int'l Journal Computer Vision, Vol. 24, No. 2, Netherlands, p137-54, 1997.

[30] Yu, T., Zhang, Y. ”Retrieval of Video Clips Using Global Motion Information,” Electronic Letters, Vol. 37, No. 14, p893-895, July 5, 2001.

Video Registration: Key Challenges & Impacts

Harpreet S. Sawhney
Sarnoff Corporation
CN5300
Princeton, NJ 08543
hsawhney@sarnoff.com

1. Introduction

Digital video is increasingly becoming a form of data and information that can be viewed, exchanged, manipulated, abstracted, archived and retrieved. In digital form, video becomes a source of information about the world independent of the observer, in contrast with traditional analog video whose lines and frames carry meaning only through viewing. The information content in videos consists of three forms: spatial geometry and locations, temporal trajectories and events, and appearance of objects and scenes. All the three forms of information are potentially extractable from videos through the exploitation of the temporal aspect of videos as opposed to static still imagery. Temporal exploitation of videos can assist in exposing the spatial and temporal coherence of scene surfaces and objects and in turn can lead to the extraction of geometry, trajectory and appearance attributes. Video registration is an important tool in the spatio-temporal exploitation of digital video. Key advances in algorithms for video registration and sustained increase in the computational prowess of standard and custom platforms has taken video registration into the realm of real world applications as a single most deployed computer vision technology. This paper highlights the key challenges in video registration and how overcoming some of these has impacted the real world.

2. A Framework for Video Registration

Video registration is the process of aligning ensembles of pixels over multiple time instants with the goal of exposing and representing the underlying spatio-temporal behavior of scenes and objects. Algorithms and applications for video registration can be categorized based on how they represent and extract the following three entities.

1 Models of motion or transformation and structure between video frames.

2 Domains of validity of the transformations, that is, surface and object masks or shapes.

3 Models of appearance of surfaces and objects.

Representation of videos in terms of these categories is generically called a *layered representation.* I will show that issues in most video registration algorithms and applications can be understood in terms of a layered representation.

3. Global Frame-to-frame Registration

When a single model of motion and optionally structure is applicable between frames of video, globally consistent registration can be done. In this case, the goal is to align video frames without extracting individual object hypothesis. A number of applications employ such global registration, e.g. mosaics, ego-motion estimation, stereo disparity estimation etc. The models of motion and structure can vary from global parametric, to global-plus-local parametric to purely local parametric [3]. I will use some of these applications to answer the question:

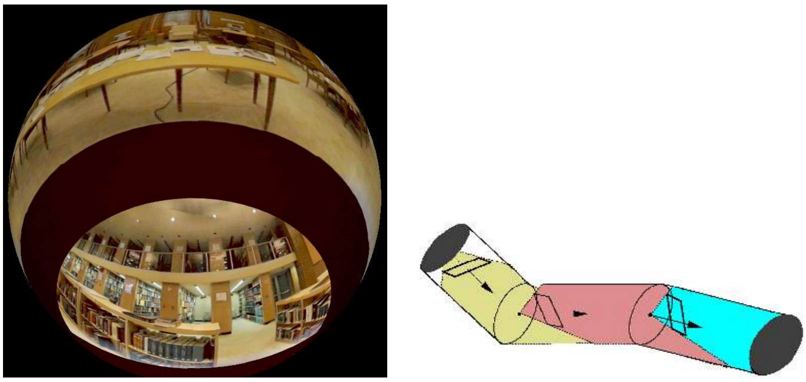


Figure 1.1. **Left:** Spherical mosaic computed iteratively using frame-to-frame alignment, topology inference and global consistency. **Right:** A schematic showing the idea of 1D manifold mosaicing by creating piecewise linear pipe mosaics. (*Image courtesy Shmuel Peleg.*)

What has been accomplished in video frame-to-frame registration in the context of mosaics and panoramas, and what are the new challenges ?

Frame-to-frame video registration for mosaics and panoramas is largely considered a solved problem. There are some novel aspects of mosaic creation that have been highlighted in the ”2D mosaicing” work of [10] and manifold mosaicing work of [9]. The 2D mosaicing work highlights how frame-to-frame registration can be used in an iterative algorithm to infer the spatial arrangement of frames (2D topology) as well as to create local constraints that can be combined to create a globally consistent mosaic. Fig. 1.1(left) shows an example of a spherical mosaic created from a video of about 600 frames using frame-to-frame alignment, topology inference and global consistency. A novel concept based on frame-to-frame video alignment was introduced by Rousso et al. [9] in the form of universal pipeline mosaics or 1D manifold mosaics. A schematic of the idea is shown on the right panel in Fig. 1.1.

Frame-to-frame alignment in video frames typically does not have to cope up with large appearance changes. However, alignment between multi-modal sensors, for example visible and IR, requires matching algorithms that can deal with large appearance changes in addition to solving for global geometric transformations. Recent work has combined global geometric alignment with correlation based iterative matching and invariant feature representations. This answers the following question raised by the panel organizers:

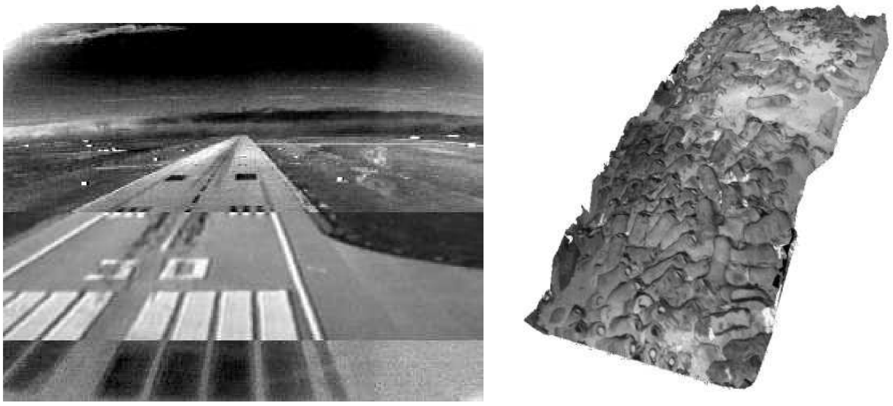


Figure 1.2. **Left:** Correlation based 2D affine parametric alignment between an IR frame and a visible spectrum video frame, shown as split window. The top half shows the IR frame and the bottom half shows the visible frame. (*Image courtesy Michal Irani.*) **Right:** Correlation based direct computation of depth using an underwater video of Greek amphoras on the ocean floor from a Roman shipwreck [7].

Does correlation still play an important role in registration? Has any thing new happened in correlation during the last 50 years?

Correlation plays an important role in registration in situations where there may be large appearance changes between frames. Irani and Anandan [5] proposed a novel role for correlation in image registration and presented correlation based alignment of multi-modal images in an iterative parameter estimation algorithm. They used oriented energy pyramids as input features over which iterative correlation based alignment is done. Fig. 1.2(left) shows an example of alignment of a video image with and IR image that are related through a 2D affine transformation. The idea of correlation based parametric alignment was extended to alignment using 3D rigid motion model and local depth by Mandelbaum et al. [7]. An example of 3D reconstruction based on correlation and video alignment of an ancient Greek shipwreck on the ocean floor is shown in the right panel of Fig. 1.2. The underwater videos used to create the shape have strong appearance changes due to the use of a ”miner’s lamp” illumination for video capture.

Video Alignment with 3D/Stereo Constraints

One of the important applications of image and video alignment is in the computation of 3D depth maps from rigid motion and stereo constraints. The key problems in depth estimation are: (i) obtaining depth maps with sharp depth discontinuities, (ii) reliable estimation in textureless regions, (iii) recovering depth of thin structures, and (iv) reasonable estimation for unmatched areas. The past few years have seen progress in all these problems [4, 15]. However, dealing with all the problems within a single algorithm with a real-time implementation is still an outstanding challenge. Furthermore, in the context of dynamic synchronized multi-camera capture for virtualized reality [8], high quality dynamic depth estimation is still in its infancy [16, 14]. In virtualized reality and image based rendering applications, instead of the absolute accuracy of the depth map, a measure of performance is the quality and speed of generation of novel views [12]. Dynamic depth extraction needs to exploit both 3D spatial and temporal constraints in video alignment.

The Role of Pose Data & Features in Video Alignment

Precise alignment of current video data of a locale with stored reference imagery (called geo-registration for alignment to geo-imagery) has key applications in change detection, targeting and visualization. Typically the reference imagery also consists of digital elevation maps (DEMs) and video platforms provide approximate pose information in terms of Engineering Support Data (ESD) and telemetry. However, the error in video to reference alignment can be as high as 200-500m. with dead reckoning using the pose data. Therefore, the challenge is to use image based alignment techniques to improve the alignment accuracy to the sub-meter range. Since typically the video and reference imagery differ tremendously in their appearances, use of multi-resolution features for coarse matching is important. Furthermore, fine scale alignment involves aligning not each individual frame of video independently to the reference, but using video frame-to-frame constraints over a window for bundle-of-frame alignment. Tremendous progress towards a robust real-time system for this problem has been reported in [18]. A sample result from the system is shown in Fig. 1.3. A challenging problem in geo-registration is that of coarse indexing when the pose data may not be available or may be inaccurate by 1000's of pixels. Promising work in this direction has been presented in [11]. Coarse indexing and fine alignment in geo-registration become especially challenging when the video platforms use highly zoomed cameras with fractional fields of view to capture high resolution data from large distances.

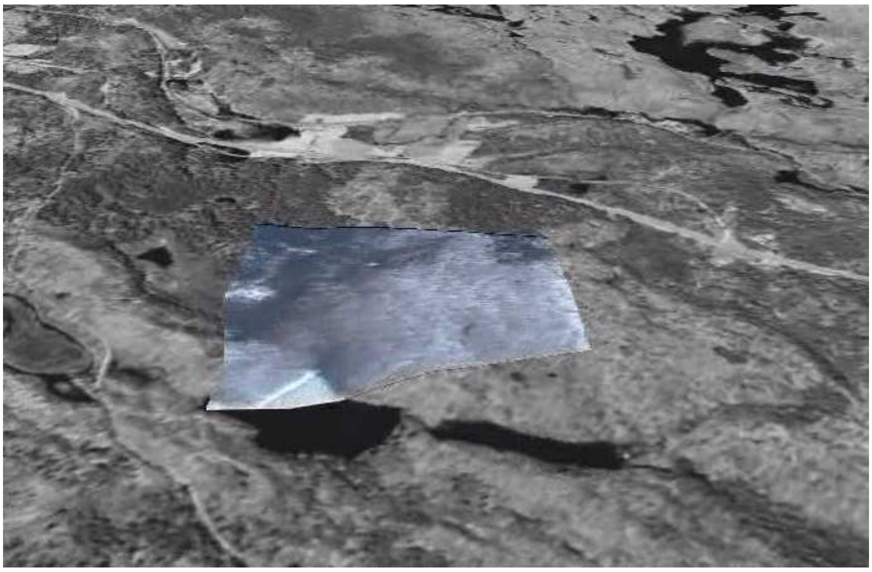


Figure 1.3. A video of snow cover aligned with reference imagery with foliage and large terrain variation.

4. Estimation of Object & Surface Masks

The second important ingredient of video registration problems is the need to extract object and surface masks in the process of registration. Many applications would benefit from the extraction of precise object masks. Shape coding in MPEG4, and rotoscoping in entertainment applications are two examples. Furthermore, robust tracking of multiple objects in clutter requires representation and extraction of objects in videos. Although much progress has been made in the representation and extraction of objects from videos, demands on precision still require more work. Wang & Adelson [17] introduced representation and extraction of object and surface layers from video. The machinery provided by mixture models and the Expectation-Maximization (EM) algorithm provides a nice computational framework for solving the parametric layer estimation problem [1]. Initial work on extending layer representations to 3D layers with plane plus parallax is presented in [2].

5. Models of Object Appearance & Shape

Video registration can be employed to extract a complete representation of objects : their motion, appearance and shape. A number of applications require such a complete representation. Object tracking is one capability that can be used to create partial or complete representations of objects for applications in security and surveillance, biometrics and indexing. Numerous research efforts have been devoted to the problem of tracking and video based object extraction. I will highlight works in which all the three ingredients of an object representation have been used. An extension of extraction of object layers for dynamic tracking in aerial video surveillance was presented in [13]. This work combined simple 2D shape models with 2D motion and appearance model as 2D layers that are tracked as a complete state over time. In [6], a simplified 3D human body model was combined with learned appearances of foreground and background to perform multi-object tracking within a Bayesian particle filtering framework.

When tracking is posed as the maintenance of identity of an object through clutter and in the presence of other objects, it becomes a problem of video registration in which a complete object representation needs to be employed. A complete object and background representation becomes necessary if real-world situations need to be handled, for instance, multiple objects crossing each other, objects coming to a stop and then moving again, similar objects moving alongside each other etc.

6. Successes & the Future

A number of real world commercial applications deploy video registration as a core piece of technology. It may not be an overstatement to say that video registration is the single most deployed computer vision technology. Princeton Video Imaging (PVI) uses pattern indexing and registration to locate patches in the video where electronic advertisements are inserted in broadcast video. Video mosaicing has been commercialized by a number of companies: Live Picture, VideoBrush (now IPIX), RealVIZ to name a few. *MatchMove* is another technology that uses video registration for establishing correspondences as a step towards solution of camera poses. This technology is being widely commercialized by companies like RealVIZ, 2D3 and Synapix. The US military has funded work on video geo-registration and the promise of recent results may bring it to the status of a deployed technology soon.

A number of challenges in algorithms and applications for video registration lie ahead. With the increased need for deployable security and surveillance technologies, it is expected that advances will be required and will happen in a number of areas requiring various aspects of video registration: (i) robust tracking in crowded scenarios, (ii) increasing resolution of faces and human forms from low quality security videos, (iii) high quality capture of biometric features (faces, irises etc.) at a distance, etc. A challenging problem that could combine online tracking and offline access is that of object indexing based on video. Robust tracking and model acquisition in real-time can be used to populate an online database. The stored models can be used later to match against new objects and to provide indexing capabilities.

On the algorithmic side, high quality extraction of depth and 3D models of objects remains a challenge. This is especially true of applications requiring high quality rendering from models and textures acquired from dynamic imagery. As video cameras and multi-camera systems become more prevalent, real-time systems for dynamic model extraction and rendering will be required. Currently without such a real-time capability, some of the deployed systems (EyeVision and Kewazinga in sports broadcast applications) have to live with systems with numerous cameras and relatively low resolution and quality output.

References

[1] S. Ayer and H. S. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and mdl encoding. In *International Conference on Computer Vision*, pages 777–785, Cambridge, June 1995.

[2] S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *CVPR98*, pages 434–441, 1998.

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

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *ICCV99*, pages 377–384, 1999.

[5] M. Irani and P. Anandan. Robust multi-sensor image alignment. In *Proc. Intl. Conf. on Computer Vision*, 1998.

[6] M. Isard and J.P. MacCormick. Bramble: A bayesian multiple-blob tracker. In *ICCV01*, pages II: 34–41, 2001.

[7] R. Mandelbaum, G. Salgian, and H. S. Sawhney. Correlation-based estimation of ego-motion and structure from motion and stereo. In *Proc. Intl. Conf. on Computer Vision*, 1999.

[8] P.J. Narayanan, P.W. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. In *ICCV*, pages 3–10, 1998.

[9] B. Rousso, S. Peleg, I. Finci, and A. Rav-Acha. Universal mosaicing using pipe projection. In *ICCV98*, pages 945–952, 1998.

[10] Harpreet S. Sawhney, Steve Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. In *ECCV*, pages 103–119, 1998.

[11] C. Schmid and R. Mohr. Combining greyvalue invariants with local constraints for object recognition. In *Proc. Computer Vision and Pattern Recognition Conference*, pages 872–877, 1996.

[12] R. Szeliski. Prediction error as a quality metric for motion and stereo. In *Proc. Intl. Conf. on Computer Vision*, 1999.

[13] H. Tao, H.S. Sawhney, and R. Kumar. Dynamic layer representation with applications to tracking. In *CVPR00*, pages II:134–141, 2000.

[14] H. Tao, H.S. Sawhney, and R. Kumar. Dynamic depth recovery from multiple synchronized video streams. In *CVPR01*, 2001.

[15] H. Tao, H.S. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *ICCV01*, pages I: 532–539, 2001.

[16] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *ICCV*, pages 722–729, 1999.

[17] J. Y. A Wang and E. H. Adelson. Layered representation for motion analysis. In *Proc. Computer Vision and Pattern Recognition Conference*, pages 361–366, 1993.

[18] R.P. Wildes, D.J. Hirvonen, et al. Video georegistration: Algorithm and quantitative evaluation. In *ICCV01*, pages II: 343–350, 2001.

Video Registration: Key Challenges

Richard Szeliski
Microsoft Research One Microsoft Way
Redmond, WA 98052
szeliski@microsoft.com

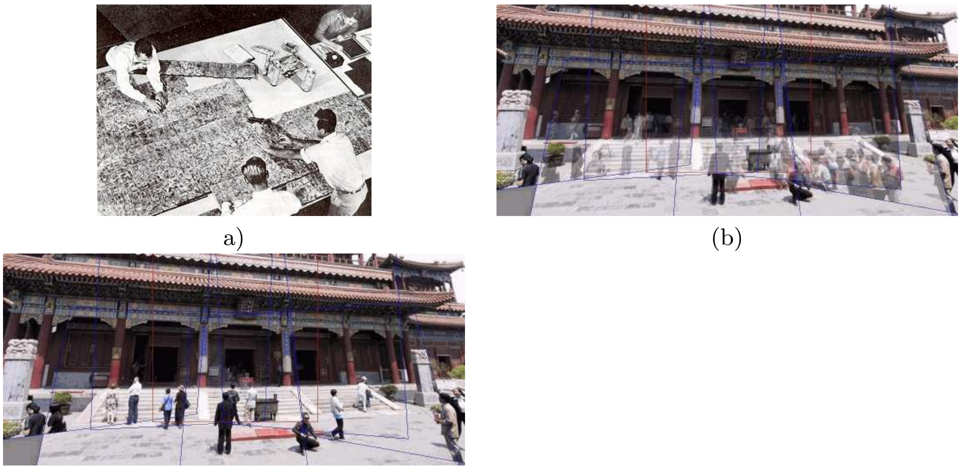
1. Introduction

Video registration today plays an important role in a number of applications. These include motion estimation for video compression, mosaic creation and change detection for consumer photography and surveillance, and the visual overlay of advertizing and positional information (e.g., 1st down markers) onto live video. These applications are all possible because of fundamental breakthroughs in video registration made over the last decade.

However, many other applications await the solution of even more complex problems. These problems include:

- the creation of seamless mosaics that compensate for exposure differences and foreground object movement;
- pixel-accurate correspondence algorithms that correctly account for (semi-)occluded regions and textureless regions;
- dealing with transparent and specular motion as separate layers;
- dealing with non-rigid and repetitive motions.

In the following sections, I describe in more detail these open problems and some potential solutions.



a)

(b)

(c)

Figure 1.1. Mosaics: (a) early photomosaics; (b–c) before and after de-ghosting.

2. Mosaics: dealing with ghosts

The automatic construction of mosaics has progressed rapidly over the past decade. Early photomosaics were created by photogrammetrists using manual techniques (Figure 1.1a), and the later using computers [16]. Fully automated techniques for scene stabilization, video summarization, change detection, and object insertion were developed at Sarnoff labs in the 1990s [10, 11].

Automated mosaic construction for consumer applications began in earnest with the introduction of Apple's QuickTime VR (QTVR) [8], and was later extended to handle arbitrary camera motion using perspective and rotational motion models [22, 20]. A good survey on panoramic mosaicing techniques can be found in [2].

One of the open problems in mosaic construction is how to deal with moving objects and with parallax induced by non-rotational camera motion, both of which may result in "ghost" images. When the moving objects are small and the image sequence is dense, median filtering can be used succesfully [11]. If the scene is static, the parallax can often be computed using stereo matching techniques and then removed [15].

When only a small amount of overlap exists, the ghosts can be eliminated by carefully cutting image seams at locations of least error [17, 9]. However, when a large number of overlaps exist, the strategy of finding pairwise optimal seams does not work. We have recently developed a technique that finds all potential ghosted regions (by measuring the local variance in the contributing pixels at each location), and then de-

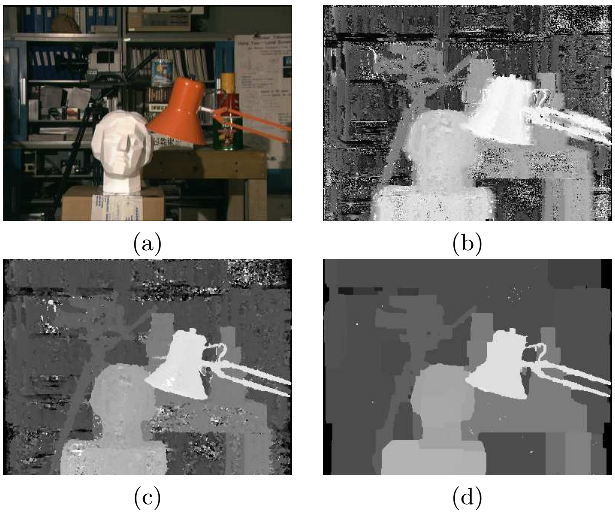


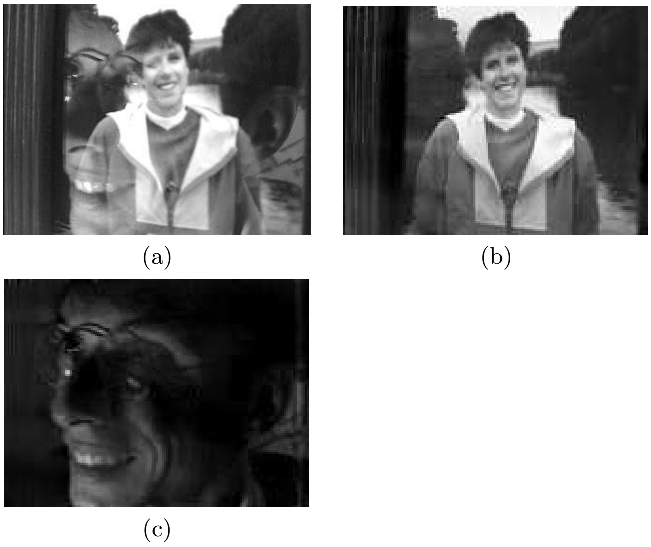
Figure 1.2. Accurate stereo correspondence: (a) sample image from sequence (courtesy U. Tsukuba); (b) typical 5×5 window-based correspondence; (c) using spatially and temporally shiftable windows; (d) after applying graph-cut optimization. Note how the motion boundaries are much more accurate.

cides which of the regions to keep and which to throw away [26]. Our decision algorithm is based on a weighted vertex cover, which ensures that at most one image can contribute to any potentially ghosted region. Our algorithm also compensates for exposure differences using a local spline-based exposure compensation mechanism. A result of using our algorithm can be seen in Figure 1.1.

While these results are encouraging, there is more work to be done. For consumer-level stitching, fully automated techniques that can deal with little overlap and arbitrary image ordering need to be developed. As well, better de-ghosting that can "guess" which objects to keep and which to leave (e.g., not to leave in two copies of a given person) would be useful.

3. Correspondence: occlusions and un-textured regions

Correspondence algorithms are at the heart of motion estimation, optic flow, and stereo matching algorithms. A tremendous amount of work has been done in making these algorithms efficient and accurate [3, 1, 18]. However, most algorithms still have trouble in regions that are partially occluded and in textureless regions.



(a) (b)

(c)

Figure 1.3. Transparent motion recovery: (a) sample image from sequence (courtesy Michael Black); (b–c) recovered transparent layers.

Many different methods have been developed to deal with occlusions. Robust matching can sometimes help [5]. Dynamic programming, which is usually restricted to stereo matching with a horizotal epipolar geometry, can explicitly deal with occlusions [6]. When multiple images are available (e.g., in video streams), using spatially and temporally shifted windows can do a good job near discontinuities [14].

For untextured regions, using global optimization such as dynamic programming [6] or Markov Random Fields [7] can effectively set textureless regions to the same motion/disparity, although the DP algorithms only do so one scanline at a time. Another approach that shows much promise is to use color image segmentation in conjunction with stereo [25]. Figure 1.2 shows the results of traditional window-based stereo matching, along with our shiftable window algorithm, followed by a graph-cut MRF minimization. You can see how the overall reconstruction and especially the depth discontinuities are much more accurate using our approach.

While these recent approaches show promise, obtaining accurate correspondences near discontinuities and in textureless regions remains a challenging problem.

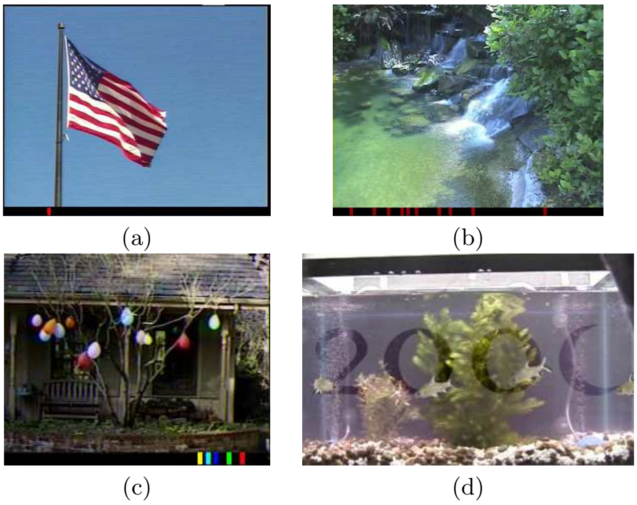


Figure 1.4. Video Texture synthesis examples: (a) waving flag, (b) waterfall, (c) tree with balloons, and (d) fish tank.

4. Transparent motion and mixed pixels

As we start applying video registration to more realistic and complex visual scenes, we need to deal with transparent and specular motion. These kinds of mixed-up multiple motions often occur because of reflections in windows or off other glossy surfaces. While several approaches have been developed to estimate the motions in such regions (e.g., [12, 4, 13], relatively little work has been done on recovering the actual images (layers) contributing to the final image. The approach we developed recently relies on dominant motion estimation followed by a least squared regression to estimate the component layers [23] (Figure 1.3). Currently, we are extending the approach to include per-pixel disparity and motion estimates.

A related problem that occurs at object boundaries is the presence of *mixed pixels*, which are colored with mixtures of foreground and background colors. While blue-screen matting techniques commonly extract the true colors and fractional opacities of such boundary pixels [21], most stereo correspondence and motion estimation algorithms do not (but see [24] for some preliminary work). In future work, we plan to study how our transparent motion model could be extended to reliably estimate fractional opacities and foreground/background colors at mixed pixels.

5. Non-rigid motion

The last open problem in video registration I would like to highlight is dealing with non-rigid motion. While flow techniques can sometimes estimate the motion in complex situations such as waterfalls, waving trees, and flames (Figure 1.4), the quality of the analysis is inadeqate if *video synthesis* is the goal. In recent work, we have developed an approach that synthesizes novel sequences with minimal explicit correspondence or motion analysis [19]. Inspired by texture synthesis approaches that copy portions of texture patches into a novel image, we paste together randomly selected subsequences of the original video and use a simple form of flow-based automated *morphing* to smooth over visual discontinuities. Even more recent approaches to this problem have used local statistical analysis techniques to describe such motions and to synthesize realistic video textures.

Video textures are just one example of a larger class of techniques that I call *video-based rendering* [19], in which source video footage is used to create synthetic animations that preserve a photorealistic look and feel. As video registration techniques continue to improve, we can expect to see more creative uses of video analysis and synthesis, which will power a new generation of important video applications.

References

[1] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, January 1994.

[2] R. Benosman and S. B. Kang, editors. *Panoramic Vision: Sensors, Theory, and Applications*, New York, 2001. Springer.

[3] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Second European Conference on Computer Vision (ECCV'92)*, pages 237–252, Santa Margherita Liguere, Italy, May 1992. Springer-Verlag.

[4] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[5] M. J. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19(1):57–91, 1996.

[6] A. F. Bobick and S. S. Intille. Large occlusion stereo. *International Journal of Computer Vision*, 33(3):181–200, September 1999.

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, November 2001.

[8] S. E. Chen. QuickTime VR – an image-based approach to virtual environment navigation. *Computer Graphics (SIGGRAPH'95)*, pages 29–38, August 1995.

[9] J. Davis. Mosaics of scenes with moving objects. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'98)*, pages 354–360, Santa Barbara, June 1998.

[10] M. Hansen, P. Anandan, K. Dana, G. van der Wal, and P. Burt. Real-time scene stabilization and mosaic construction. In *IEEE Workshop on Applications of Computer Vision (WACV'94)*, pages 54–62, Sarasota, December 1994. IEEE Computer Society.

[11] M. Irani and P. Anandan. Video indexing based on mosaic representations. *Proceedings of the IEEE*, 86(5):905–921, May 1998.

[12] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *International Journal of Computer Vision*, 12(1):5–16, January 1994.

[13] S. X. Ju, M. J. Black, and A. D. Jepson. Skin and bones: Multi-layer, locally affine, optical flow and regularization with transparency. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'96)*, pages 307–314, San Francisco, June 1996.

[14] S. B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2001)*, volume I, pages 103–110, Kauai, Hawaii, December 2001.

[15] R. Kumar, P. Anandan, M. Irani, J. Bergen, and K. Hanna. Representation of scenes from collections of images. In *IEEE Workshop on Representations of Visual Scenes*, pages 10–17, Cambridge, Massachusetts, June 1995.

[16] D. L. Milgram. Computer methods for creating photomosaics. *IEEE Transactions on Computers*, C-24(11):1113–1119, November 1975.

[17] D. L. Milgram. Adaptive techniques for photomosaicking. *IEEE Transactions on Computers*, C-26(11):1175–1180, November 1977.

[18] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7–42, May 2002.

[19] A. Schödl, R. Szeliski, D. H. Salesin, and I. Essa. Video textures. In *Computer Graphics (SIGGRAPH'2000) Proceedings*, pages 489–498, New Orleans, July 2000. ACM SIGGRAPH.

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

[21] A. R. Smith and J. F. Blinn. Blue screen matting. In *Computer Graphics Proceedings, Annual Conference Series*, pages 259–268, Proc. SIGGRAPH'96 (New Orleans), August 1996. ACM SIGGRAPH.

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

[23] R. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2000)*, volume 1, pages 246–253, Hilton Head Island, June 2000.

[24] R. Szeliski and P. Golland. Stereo matching with transparency and matting. *International Journal of Computer Vision*, 32(1):45–61, August 1999. Special Issue for Marr Prize papers.

[25] H. Tao, H.S. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *Eighth International Conference on Computer Vision (ICCV 2001)*, volume I, pages 532–539, Vancouver, Canada, July 2001.

[26] M. Uyttendaele, A. Eden, and R. Szeliski. Eliminating ghosting and exposure artifacts in image mosaics. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2001)*, volume II, pages 509–516, Kauai, Hawaii, December 2001.