ALIGNING 'DISSIMILAR' IMAGES DIRECTLY

Yaser Sheikh and Mubarak Shah

School of Computer Science
University of Central Florida
Orlando Florida, USA

ABSTRACT

This paper introduces a hierarchical algorithm for the registration of corresponding images that do not necessarily have strong global similarity, such as multi-modal images, images with varying illumination (or specular reflection) and images with significant local motion. The method is based on the global maximization of an average local correlation measure, without the generation of similarity surfaces. Fisher's Z-Transformation is used to rectify the correlation coefficient to ensure that additivity between correlation samples is strictly accurate. As a result, the proposed method can handle sizable misalignments in rotations, scale and shear. Direct error functions are also robustified to optimize alignment in the presence of outliers. The result is a completely autonomous system that recovers the global transformation between two images despite substantial visual differences, including contrast reversals, local motion and disjoint image features. The algorithm was successfully tested with a wide variety of images, particularly where conventional frame-to-frame alignment algorithms had failed.

1. INTRODUCTION

When two different sensors observe the same scene, dissimilarities are quite likely to arise between the images captured. Depending on the degree of difference between the sensors, the dissimilarities may be negligible other than a spatial transformation, as in the case of human stereo vision, or they may be pronounced, as in the case of multi-modal measurements. A specific example can be observed in Infrared and Electro-Optical images, where measurements operate in different wavelength bands, and as a result features detected in one sensor image may be absent in the other (disjoint features). Dissimilarity is also often observed in images taken of the same scene after a large interval of time. An example of this occurs in the geo-registration of aerial video images with previously recorded reference imagery. In such a case, artifacts that change or are absent may exacerbate dissimilarity already existent due to photometric differences. Even within frame-to-frame alignment of images acquired by the same sensor, the presence of local motion or specular reflection can cause dissimilarity between two consecutive frames. Of course, dissimilarity may distort the image beyond recognition, but the premise of aligning two dissimilar images in this paper is that enough correlation exists between the two images, locally *or* globally, to allow alignment between them. Aligning images in the presence of local motion and/or illumination variations has been recognized as an important problem in computer vision research and while algorithms have been developed to robustly align two similar images, existing approaches fail in aligning two images thus defined to be dissimilar. These algorithms usually make assumptions of brightness constancy or of strong global similarity - assumptions often severely violated in the situations already described. The global relationship between intensity values of corresponding pixels of two dissimilar images is usually complex and may even be non-existent, caused by phenomena such as local contrast reversal, disjoint features or multiple-intensity to single-intensity projection. Evidently, methods are needed that rely on a 'stronger' similarity metric, and that have an effective means of suppressing dissimilarities and emphasizing similarities. Furthermore, handling outliers is a central problem in such matching. Aligning despite local motion, in particular, can be thought of simply as a problem of handling outliers robustly. Since locally moving objects can be considered as a set of pixels with outlier motion, if outliers are handled appropriately, images can be aligned despite local motion. Thus, the challenges for aligning dissimilar images can be summarized as being threefold: (1) finding effective measures of similarity, (2) using a representation that suppresses dissimilarities and emphasizes similarities, and (3) employing an effective method for handling outliers. We formulate the problem as follows: For each image I_1 there are allowable transforms $T(\mathbf{x};\vec{a})$, where $\vec{a}$ denotes the model parameters of global motion and $\mathbf{x} = (x, y)$ are the pixel coordinate of the image. Parametric alignment is a search over the transformation parameters, $\vec{a}$, that maximizes some global measure of 'fit' or similarity, S, between the source image, I_1, and the goal image, I_2,

$$\max S(I_1(\mathbf{x}), I_2(T(\mathbf{x};\vec{a}))). \tag{1}$$

The remainder of the paper is structured as follows: Section 2 provides a brief examination of previous work and an overview of the proposed approach, Section 3 discusses an appropriate representation of images and reviews various models of transformations that can be used with the proposed algorithm. In Section 4, measures of similarity are discussed and compared briefly, and the choice of normalized cross correlation as a similarity metric is justified. The algorithm itself is outlined in Section 5, followed by results and the conclusion in Sections 6.

2. PREVIOUS WORK

Previous work may be broadly categorized into three approaches. The first approach handles violations to the brightness constancy through image preprocessing techniques, such as band-pass filtering or contrast gain normalization. Unfortunately, these techniques account for *specific* types of brightness constancy violations and are unable to handle complex (and unmodelled) violations. The second approach models such violations as statistical outliers, [4], [3], [19]. However, such approaches assume strong global statis-

tical similarity, and as a result are unable to address local dissimilarities like local contrast reversal and/or disjoint features, within a single framework. The third approach explicitly maximizes measures of similarity to estimate motion. In [10] an iterative algorithm is presented that uses normalized cross correlation locally on an invariant representation, generating similarity surfaces by translating a template around a radius and then searching over the sum of the similarity surfaces for the optimal parametric transformation. There are two limitations of this approach. First, the use of such similarity surfaces impedes the alignment of strong rotations or of projective foreshortening, since the translated template can only approximate small scaling and rotations as progressive translations. Second, similarity surface values (normalized cross correlation scores), are summed directly. Since the value of the correlation coefficient is not a linear function of the relationship between the images [23], it is misleading to directly sum correlation coefficients. In cases where an average or sum of correlations is required: a conversion into additive measures has to be performed. This will be explored in depth in Section 4. Other approaches address the problem by maximization of mutual information [22], an invariant similarity measure. Since this method presumes global statistical similarity between the two images, it is liable to fail where global similarity cannot be assumed. Furthermore, these methods cannot be executed in a coarse-to-fine manner and it is therefore difficult to come up with an efficient search strategy.

We propose an algorithm that uses local similarity measures, and utilizes them to directly estimate *global* similarity. However, since we do not generate similarity surfaces, our method can recover larger rotation, shear and scaling and does not degenerate when higher order parametric models of motion are used. Furthermore, we correct the correlation coefficients to validate their addition by the use of Fisher's Z-transform. Finally, we robustify the constructed error function, to allow optimal registration in the presence of outliers caused by dissimilarity in data.

3. REPRESENTATION AND MODELS

When working with dissimilar images taken from different modalities particularly, (e.g. Infra-Red (IR) images with Electro-Optical (EO) images), it is necessary to represent the images in a fashion that suppresses individual information and emphasizes joint information. Multi-sensor image pairs are correlated predominantly at higher spatial frequency [10], where the *structure* of a scene is captured. There is both biological and computational support for the use of a bank of oriented bandpass filters to extract such structural information ([6], [15]). To capture the structural information we employ a log-Gabor pyramid oriented in four directions. In order to make the representation invariant to contrast reversal, we take the absolute values of the log-Gabor response. Such a representation has the advantage of preserving directional and structural information, while remaining non-sparse. In theory any bank of oriented bandpass filters could be used, like a Laplacian Energy Pyramid (as in [10]), but it was empirically found that the Log-Gabor pyramid best emphasized the coherent structural information.

3.1. Models of Transformation

When the displacement between an image pair is small relative to the displacement from the scene or the optical center is stationary, an assumption of scene planarity holds for the situation. In such cases, where a plane can approximate the scene, a parametric transformation can model the displacement field (u, v) between two images. A taxonomy of motion models in general can be found in [2] and a detailed classification of linear and non-linear parametric transformation models can be found in [16]. The algorithm proposed in this paper can be applied to all the specified parametric motion models. Two frequently used models are the affine and projective models. The allowed set of transforms $T(\mathbf{x}; \vec{a})$ can be used to recover mapping for each coordinate,

$$T(\mathbf{x}; \vec{a}) = \begin{bmatrix} x + u(x, y) \\ y + v(x, y) \end{bmatrix} \tag{2}$$

where u and v are the optical flow vectors at each pixel coordinate. For the use of a linear transformation model, affine transformation can be estimated as

$$\begin{aligned} u(x, y) &= a_1 x + a_2 y + a_3 \\ v(x, y) &= a_4 x + a_5 y + a_6, \end{aligned} \tag{3}$$

in terms of the six affine parameters $\vec{a}$, $[a_1 \;\; a_2 \;\; a_3 \;\; a_4 \;\; a_5 \;\; a_6]$. For the general projective transformation between images, the optical flow vectors are expressed as

$$\begin{aligned} u(x, y) &= \tfrac{a_1 x + a_2 y + a_3}{a_7 x + a_8 y + 1} \\ v(x, y) &= \tfrac{a_4 x + a_5 y + a_6}{a_7 x + a_8 y + 1}, \end{aligned} \tag{4}$$

where $\vec{a}$, $[a_1 \;\; a_2 \;\; a_3 \;\; a_4 \;\; a_5 \;\; a_6 \;\; a_7 \;\; a_8]$ are the eight projective parameters. Since the proposed algorithm does not generate similarity surfaces based on translating a template, transformations involving rotation, scaling and shear do not cause degeneracy in alignment, and as a result any linear or non-linear parametric transformation can be applied in general. The algorithm has been implemented with both affine and projective models.

4. MEASURING SIMILARITY

The correlation between two variables represents the degree to which signals are related. As the problem was formulated, alignment is to be achieved by transforming the source image in such a way that a measure of similarity between the two images is maximized. A conventional measure of similarity between images is the sum of squared differences (SSD), however similarity measures based on SSD do not consider linear relationships between two images. Since photometric differences can induce large changes in illumination, it is imperative to use a measure of similarity that *does* take linear relationships into account. Although the algorithm presented here is not strictly dependent on any single similarity measure within the framework, Normalized Cross Correlation is our metric of choice since it inherently detects linear relationships. For any pair of images $I_2(\mathbf{x})$ and $I_1(\mathbf{x})$, the correlation coefficient r_{ij} between two patches centered at location (x_i, y_j) is defined as

$$r_{ij} = \frac{\sum_{w_x} \sum_{w_y} (\phi_2)(\phi_1)}{\sqrt{\sum_{w_x} \sum_{w_y} (\phi_2)^2 \sum_{w_x} \sum_{w_y} (\phi_1)^2}} \tag{5}$$

where

$$\phi_1 = I_1(\mathbf{x} + [w_x \;\; w_y]^T) - \mu_1 \tag{6}$$

$$\phi_2 = I_2(\mathbf{x} + [w_x \;\; w_y]^T) - \mu_2 \tag{7}$$

and w_x and w_y are the dimensions of the local patch around (x_i, y_j), and μ_1 and μ_2 are the patch sample means. Correlation is measured locally between patches around each pair of corresponding pixels, and this local correlation is then summed to get a *global* estimate of similarity. However, before such an addition is performed, Normalized Cross Correlation has to be converted into an *additive* measure. Additivity implies that the average of similarity coefficients in a number of samples represents an 'average correlation' in all those samples. This condition has two consequences. First, that the similarity magnitude rather than the similarity measure is used (hence avoiding a cancelling effect on summation of negative and positive measures), and second, that the similarity measure *should vary linearly with the magnitude of correlation*. This requirement is often ignored while summing correlation measures, resulting in inaccurate assessments of average similarity.

4.1. Fisher's Z-Transform

The Normalized Cross Correlation coefficient is not a linear function of the relational magnitude between the images [23], and as a result, correlation coefficients cannot simply be averaged. As a statistic, r has a sampling distribution. (If n sample pairs from two signals were correlated over and over again the resulting distribution of r scores would form a sampling distribution.) This distribution has a negative skew, which is the sampling distribution of r. It can be observed that although no value of r ever exceeds 1.0, there is a bias towards lower values of r. A transformation called Fisher's Z-transformation converts r to a value that is normally distributed and is defined as

$$z_{ij} = \frac{1}{2}(\ln(1 + |r_{ij}|) - \ln(1 - |r_{ij}|)). \tag{8}$$

As a result of having a normally distributed sampling distribution, there is an equal probability of detecting any correlation score and hence they can be meaningfully added.

4.2. Robust Estimation of Similarity

Disjoint features, local motion and photometric ambiguity may all contribute to causing outliers. It is, therefore, important to have mechanisms that ensure robustness in the presence of outliers. To that end we incorporate an M-estimator into our framework. An M-estimator minimizes an expression of the form

$$\sum_i \rho(f_i(\mathbf{x}_i, \vec{\theta}); \sigma), \tag{9}$$

where $\vec{\theta}$ are the parameters of the model being fit (in our case the transformation model parameters, $\vec{a}$), and f_i is the residual error of the model on the ith data point (in our case the corrected correlation coefficient, z_{ij}). Many ρ functions have been used for M-estimation, such as the Geman-McLure function [19], the Sigmoid function or the Sum of Absolute Difference. Through experimentation we found that a sigmoid based estimation function produced the most desirable results and consequently incorporated it into our framework. The sigmoid function is

$$\rho(z) = \frac{1}{1 + \exp(-\frac{z}{c})} \tag{10}$$

where c is parameter controlling the sensitivity noise. We can thus define a final robust similarity measure in terms of the correlation coefficient as

$$\eta(\mathbf{x}_i; \vec{a}) = \frac{1}{1 + (\frac{1-|r_{ij}|}{1+|r_{ij}|})^{\frac{c}{2}}}. \tag{11}$$

5. ALIGNING THE IMAGES

For successive images acquired by a single camera, the photometric transformation is usually negligible, but for dissimilar images, particularly those captured by multi-modal sensors, this relationship is often complex and unmodelled. As a result, the two images are not often globally correlated. Even statistical correlation may not be distinctively strong between two such images. However, locally, within small patches containing corresponding information, statistical correlation is observed to be significantly higher, and as a result local similarity measures are powerful cues to estimate global alignment between dissimilar images. In this work, a measure of global similarity is estimated based on an average of local similarities, which is then maximized using numerical optimization.

5.1. Global Similarity Estimation

To utilize the strengths of any arbitrary similarity measure within the framework of direct registration, a global function of error is defined at each pixel (x_i, y_j) over a local neighborhood. The coefficient defined in (11) is used as the similarity measure with a projective transformation model, though extension to different models and different similarity measures is straightforward. It should be noted that correlation surfaces are *not* being calculated here. At every pixel (x_i, y_j), a similarity score, $\eta(\mathbf{x}_i; \vec{a})$ is calculated between two patches of size $w_x \times w_y$ centered around (x_i, y_j). Since the similarity score is additive, the global similarity measure $H(\vec{a})$ is then defined summing for all (i, j) as

$$H(\vec{a}) = \sum_i \eta(\mathbf{x}_i; \vec{a}) \tag{12}$$

This function is then maximized over the parameters of transformation using Quasi-Newton Optimization. This algorithm iteratively builds up curvature information to formulate a quadratic model problem. Gradient information is required, which is provided using numerical differentiation methods (finite differences). Finite differences involve perturbing each of the projective parameters, and calculating the rate of change of the objective function. In order to capture large motions efficiently and to counter the effects of local minima, the algorithm was implemented in a hierarchical fashion over a Gaussian pyramid. Typically, three major iterations are performed at each level of a four level pyramid. The steps of the algorithm may be summarized as follows:

1. For each coordinate position (i, j) calculate the local similarity $\eta(\mathbf{x}_i; \vec{a})$ between the two 5×5 cell around $I_1(\mathbf{x})$ and $I_2(T(\mathbf{x}_i; \vec{a}))$ using normalized cross correlation. Sum $\eta(\mathbf{x}_i; \vec{a})$ for all (i, j) to evaluate the global measure of similarity.
2. Calculate $\delta\vec{a}$, the update for $\vec{a}$, using Quasi-Newton Maximization of the objective function, $H(\vec{a})$.
3. Update $\vec{a}' = \delta\vec{a} \cdot \vec{a}$. Return to step one and repeat until exit condition is fulfilled.

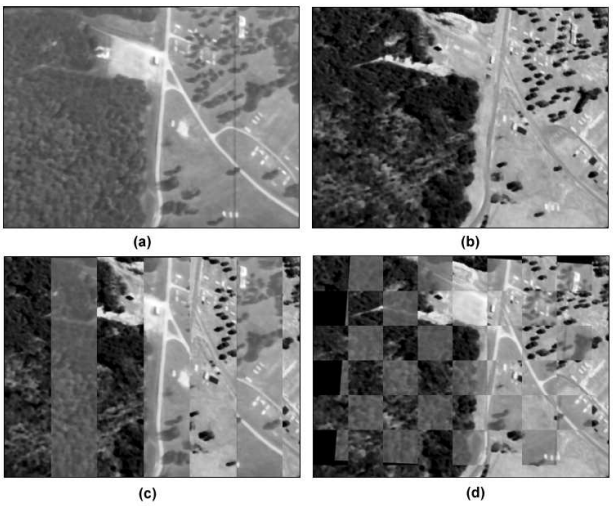

Fig. 1. Registered Aerial Image (a) Aerial Video Frame (b) Perspective-Projection of Area Reference Image (misaligned due to noisy meta data) (c) Striped display before alignment (d) Checkered Display after alignment

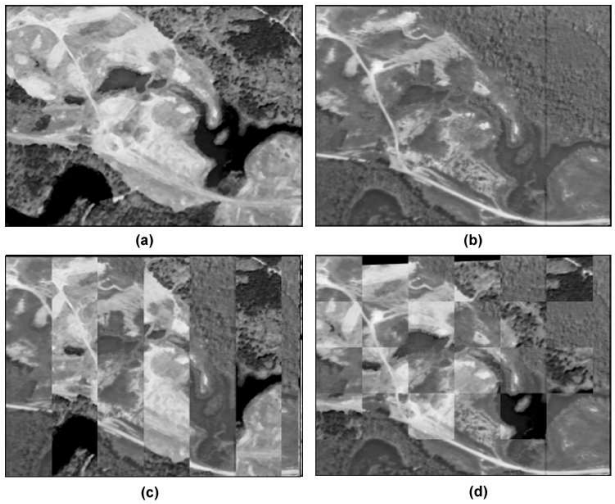

Fig. 2. Registered Aerial Image (a) Aerial Video Frame (b) Perspective-Projection of Area Reference Image (misaligned due to noisy meta data) (c) Striped display before alignment (d) Checkered Display after alignment.

Since setting an error condition on the basis of error tolerance depends on the degree of dissimilarity between two images, instead we can set an upper limit on the number of function evaluations or whether the magnitude of $\delta\vec{a}$ is below a threshold as an exit condition.

6. RESULTS AND CONCLUSION

The proposed algorithm was tested on a diverse set of images, since dissimilarity may manifest itself in a variety of ways. The context of the first set of experimentation was the alignment of aerial video images with a reference image (perspectively projected using noisy meta-data). Since there is a sizable lapse in time between the capture of the two images, and due to the intrinsic differences in the sensors, there are large visual differences between the two images, including contrast reversal and disjoint features. Despite the substantial illumination change to the extent of contrast reversal, examination of the results shows a precise pixel-wise alignment. Quantitative analysis was carried out on two aerial video clips. On the first clip, a pre-registration average error of 47.68 pixels with a standard deviation of 12.47 and a post-registration average error of 4.34 pixels and standard deviation of 3.19 per frame was recorded. On the second clip, a pre-registration error of 51.43 pixels with a standard deviation of 14.66 and a post-registration average error of 3.46 pixels and a standard deviation of 2.91 was recorded. As ground truth was not available to assess the error automatically, manual measurement was performed per frame. Figure 1 and Figure 2 show an example of the initial Video Frame and Reference Imagery before and after registration. In Figure 1, significant rotation was compensated for despite the degree of visual difference. Results are far more dramatic when seen dynamically overlaid.

The second set of experimentation tackled was the alignment of multi-modal images. The successful registration of IR images with EO images shown in Figure 3. In this type of imagery contrast reversal and disjoint features are quite common. This data set tested the algorithm's capability of handling contrast reversals, and outliers caused by the disjoint features. Notice that a large translation and scaling was recovered successfully. Fig 4 are also shown for the registration of a CT-Slice with an MR-Slice Image. A rotation of almost 40° was compensated for.

The third set of experimentation involved testing the method for alignment despite local motion. Figure 5 shows a pair of mosaics constructed despite large local motion in an opposite direction to the ego-motion of the camera. Alignment for the sequence failed using the frame-to-frame alignment algorithm of [16].

The method was tested successfully for two different motion models (affine and projective) and tested with normalized cross correlation as a measure of similarity. A high degree of accuracy was observed, despite the challenging and diverse data-set. As an implementation detail, it should be noted that while using Normalized Cross Correlation a special case needs to be handled: uniform patches, where all pixels have equal intensity. Since such patches contain no information and cause a divide-by-zero error: the similarity measure for such a case is treated as nil.

In conclusion, an algorithm has been proposed that aligns two images despite visual dissimilarities that may exist between them. In the absence of strong global similarity, an average of local similarity is maximized using normalized cross correlation to recover alignment parameters. The notion of considering pixel identities rather than pixel intensities allows direct alignment despite outlier-inducing characteristics like disjoint information, local motion, and photometric distortion etc. Future direction of this work includes the use of multi-frame information to improve similarity measurements.

7. REFERENCES

[1] P. Anandan, *"A computational framework and an algorithm for the measurement of visual motion "*, International Journal

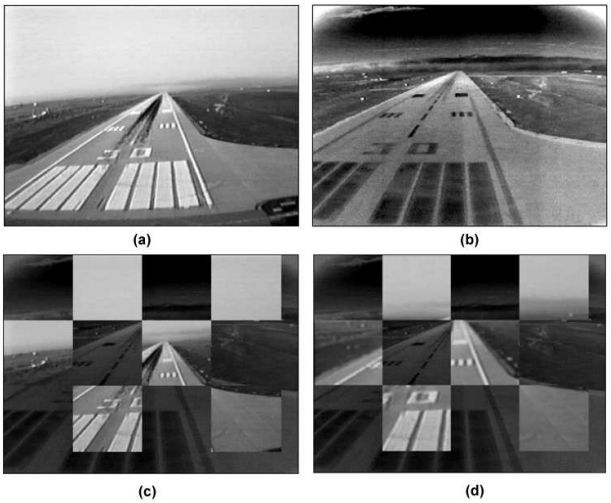

Fig. 3. Alignment between IR and EO images (a) EO Image (b) IR Image (c) Checkered Display before alignment (d) Checkered Display after alignment.

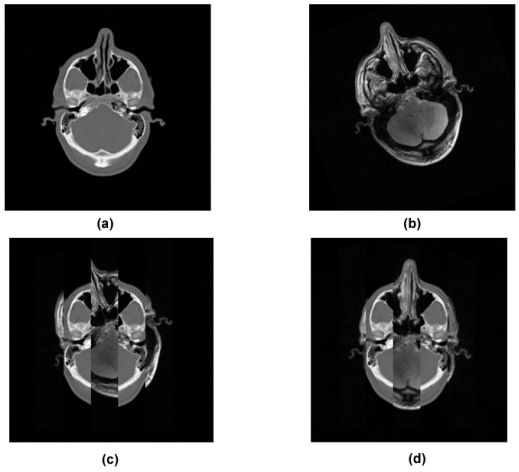

Fig. 4. Alignment between CT and MR Slices (a) CT Slice (b) MR Slice (c) Striped Display before alignment (d) Striped Display after alignment

of Computer Vision, vol.2, pp. 283-310, 1989.

[2] J.Bergen, P. Anandan, K. Hanna, R. Hingorani, *"Hierarchical model-based motion estimation"*, Proc. European Conference on Computer Vision, pp. 237-252, 1992.

[3] M. Black, and P. Anandan, *"A framework for the robust estimation of optical flow"*. International Conference on Computer Vision, pp. 231 -236, 1993.

[4] M. Black, D. Fleet, and Y. Yacoob, *"Robustly estimating changes in image appearance"*, Computer Vision and Image Understanding, 78(1):8–31, 2000.

[5] L. Brown, *"A Survey of Image Registration Techniques"*, ACM Computing Surveys, 24(4), pp. 325-376, 1992.

[6] D.H. Field, *"Relations between the statistics of natural images and the response properties of cortical cells"*, JOS A, vol 4, pp. 2379-2394, 1987.

[7] J. Foley, A. van Dam, S. Feiner, J. Highes, *"Computer Graphics, Principles and Practices"*, Addison-Wesley, 1990.

[8] K. Hanna, H. Sawhney, R. Kumar, Y. Guo, S. Samarasekara, *"Annotation of video by alignment to reference imagery"*, IEEE International Conference on Multimedia Computing and Systems, vol.1, pp. 38 - 43, 1999.

[9] B. Horn, B. Schunk, *"Determining Optical Flow"*, Artificial Intelligence, vol. 17, pp. 185-203, 1981.

[10] M. Irani, P. Anandan, *"Robust Multi-Sensor Image Alignment"*, International Conference on Computer Vision, 1998.

[11] M. Irani and P. Anandan, *"Video Indexing Based on Mosaic Representations"*. Proceedings of IEEE, 1998.

[12] B. Kamgar-Parsi, J. Jones, A. Rosenfeld, *"Registration of multiple overlapping range images: scenes without distinctive features"*, Computer Vision and Pattern Recognition, pp. 282-290, 1989.

[13] R. Kumar, H. Sawhney, J. Asmuth, A. Pope, S. Hsu, *"Registration of video to geo-referenced imagery"*, Fourteenth International Conference on Pattern Recognition, vol. 2. pp.1393-1400, 1998.

[14] B. Lucas and T. Kanade. *"An iterative image registration technique with an application to stereo vision"*, Proceedings of the 7th International Joint Conference on Artificial Intelligence, pp. 674-679, 1981.

[15] J.Malik and P. Perona, *"Preattentive texture discrimination with early vision mechanisms"*, JOS A, vol. 7, pp. 923-932, 1990.

[16] S. Mann, R.W. Picard, *"Video orbits of the projective group: a simple approach to featureless estimation of parameters"*, IEEE Transactions on Image Processing, 6(9) , pp. 1281 - 1295, 1997.

[17] J. Nocedal, S. Wright, *"Numerical Optimization"*, Springer-Verlag, 1999.

[18] J. Rodriquez, J. Aggarwal, *"Matching Aerial Images to 3D terrain maps"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(12), pp. 1138-1149, 1990.

[19] H. Sawhney, S. Ayer, *"Compact representation of videos through dominant and multiple motion estimation"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(8):814-831, 1996.

[20] R. Szeliski, H. Shum, *"Creating Full View Panoramic Image Mosaics and Environment Maps"*, Computer Graphics Proceedings, SIGGRAPH, pp. 252-258, 1997.

[21] R. Szeliski, *"Image mosaicing for tele-reality applications"*, IEEE Workshop on Applications of Computer Vision, pp. 44-53, 1994.

[22] P. Viola and W. M. Wells, *"Alignment by maximization of mutual information."*, International Journal of Computer Vision, 24(2) pp. 134-154, 1997.

[23] J.K. Wani, *"Probability and Statistical Inference"*, Appleton-Century-Crofts, New York, 1971.

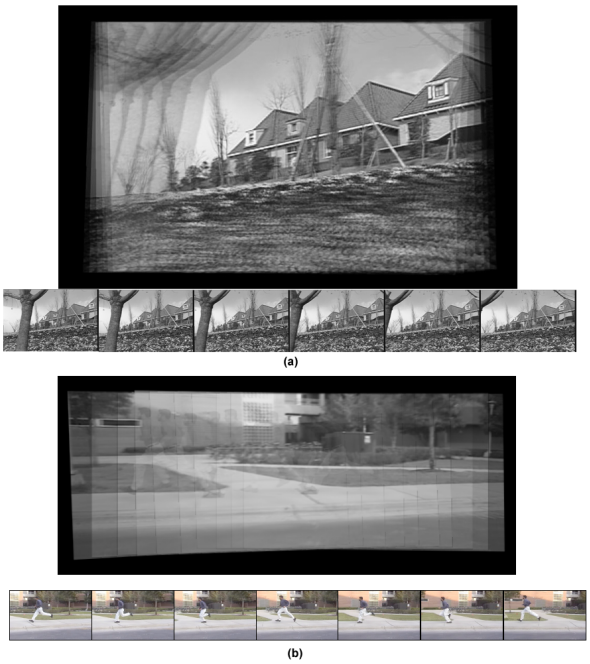

Fig. 5. Registration of Images with local motion. The mosaic was constructed by simply averaging each pixel value at a location. Ghosts were intentionally maintained to illustrate the local motion. (a) Flower Garden Sequence: Considerable parallax is observed in this sequence, yet tight dominant motion recovery is achieved. We did not use consecutive frames; the constituent frames are appended below mosaic. (b) Projective Image alignment of nine frames despite large local motion. Despite significant local motion, exact alignment is achieved. It should be noticed that the local motion is diametrically opposed to the global motion.

[24] R. Wildes, D. Hirvonen, S. Hsu, R. Kumar, W. Lehman, B. Matei, W.-Y. Zhao *"Video Registration: Algorithm and quantitative evaluation"*, Proc. International Conference on Computer Vision, Vol. 2, pp. 343 -350, 2001.

[25] Q. Zheng., R. Chellappa. *"A computational vision approach to image registration"*, IEEE Transactions on Image Processing, 2(3), pp. 311 -326, 1993.