Coupled, Multi-resolution Stereo and Motion Analysis

Chandra Kambhamettu, K. Palaniappan and A. Frederick Hasler[†]
Universities Space Research Association
[†] Lab. for Atmospheres, Code 912
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

Abstract

We propose a novel approach for fusing multi-resolution stereo and motion (both rigid and non-rigid) analysis in order to complement each other's performance. A hierarchical frame-work is presented to couple motion correspondences and stereo correspondences in order to generate accurate disparity map and motion parameters. One scenario for such system is the analysis of time-varying multi-spectral observations of clouds from meteorological satellites. Our experiments involve such time-varying remote sensing stereo data sets, and the motion is typically non-rigid as the clouds undergo shape changes. Rigid motion matching may still be performed for initial fusion, and gradually raised to non-rigid motion matching as in a coarse-to-fine strategy. Both stereo disparities and motion correspondences are estimated using such multi-resolution coarse-to-fine strategy to a sub-pixel accuracy. Experimental results using time-varying data of visible channel from two satellites in geosynchronous orbit is presented for the Hurricane Frederic.

1 Introduction

Stereo analysis of binocular images and motion analysis of monocular images is common for estimating 3D structure and motion. Both stereo and motion analysis involves solving for point correspondences, or solution for nonlinear equations, thus having an underdetermined problem on hand. Worse, when the observed scene is a time-varying imagery involving shape changes, one has to find a way to couple the problems of stereo and motion analysis with consistency checks between both the point correspondences. In this paper, we present a frame-work to couple motion correspondences and stereo correspondences in order to generate more accurate disparity map and motion parameters. Since our experiments involve time-varying remote sensing data sets, motion is typically non-rigid in our case. We also demonstrate that the frame-work is useful in general, for fusing multi-resolution stereo and motion analysis. The term "multi-resolution" is used to represent coarse-to-fine approach taken in computing disparity or motion (elaborated in later sections).

In the past, several researchers investigated fusing the stereo and motion analysis so as to compliment each other in generating accurate results. In [11], authors use the image flow fields from parallel stereo cameras to determine the relative 3D translational camera motion with respect to objects in view and to establish stereo correspondences of features in the left and right images. Optimal motion and structure estimation in presence of known and unknown noise is presented in [17] using a two-step process. First step uses a linear algorithm for a preliminary estimate, and the second step involves minimizing an optimal objective function using the previous result as an initial estimate. In [16], projective structure from un-calibrated images using an invariance relation across two or more views and the homographies of two arbitrary virtual planes is presented. In this work, authors consider a projective description of the world (object space is a 3D projective space, and image space is a 2D projective space) and develop a new shape representation, called projective depth which can be computed using a linear equation. Extraction of 3D shape from optic flow using differential invariants is considered in [3], where the ratio of principal curvatures is used as an intrinsic measure.

Most researchers in the area of structure from motion use either structure to determine motion, or motion to determine structure in order to couple the two problems. However, designing a synergistic integration of these two modules is a difficult task. We present an integrated system having these two modules (stereo and motion) which iteratively refine each other at different resolutions. The system includes both rigid and non-rigid motion analysis, and is extendable to various different motion algorithms. The structure estimation module presently has only stereo analysis, however, it can be extended to include shape-from-X modules at different resolutions. One of the important applications of such a system is in remote-sensing, where accurate cloud heights and winds are important for a host of applications such as radiation balance estimation for Mission to Planet Earth type climate baseline studies, meteorological physically-based numerical model data assimilation, cloud model verification, cloud-wind height assignment and convective intensity estimation [5]. Our main thrust in this paper is the application of the integrated system in determining cloud height and cloud wind measurements on the stereo-scopic weather satellite images. We now present the overall system of our algorithm.

0-8186-7190-4/95 $4.00 © 1995 IEEE

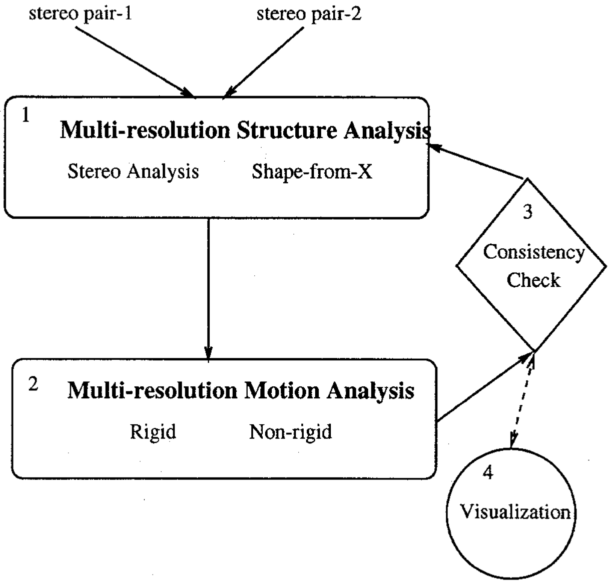

Figure 1: Integrated system

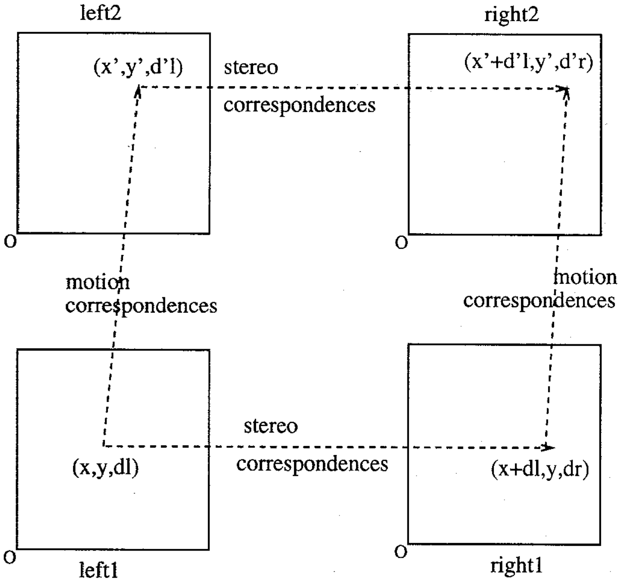

Figure 2: Consistency Check

2 Integrated system

Figure 1 indicates the overall block diagram of an integrated system for structure and motion analysis. The first block represents structure estimation module, with two sub-modules in it. In actuality, there can be many more modules of different techniques to estimate structure. They can then be used for a better estimation of disparity. Or, there can be a combination of all the techniques as an integrated system [14]. In our present implementation, we only have one module with multi-resolution stereo analysis implemented in parallel architecture (which will be explained later). The second block represents motion analysis of the estimated depth maps, provided by the first block. Third block checks for a consistency between estimated stereo correspondences (thus, estimated depth map) and motion correspondences and refines either of the correspondences if necessary. Fourth block is used to interactively visualize the results. We use an Interactive Image Spread Sheet (IISS), enhanced to suit our purpose [6].

The system is initiated with an input of 2 stereo pairs, separated by a time-lag. In our experiments, the input is either 2 GEO(Geo-synchronous Earth Orbiting)/LEO(Low Earth Orbiting) pairs or 2 GEO/GEO pairs of satellite cloud images in the visible channel, indicated by left1, right1, and left2, right2. Disparity maps are computed by searching for image correspondences between each stereo pair in either directions. i.e, disparity is generated with left image as reference and right image consisting of search space (left-right) given by (dl), and then with right image as reference and left image consisting of search space (right-left), given by (dr). Ideally, the disparity dr should be a negative of dl, indicating consistent stereo correspondences in both the directions. Any inconsistencies may be attributed to occlusion, method of stereo analysis chosen, or not enough image structure at a given point for a given resolution. Thus, the output of the first block will be 4 disparity maps: dl1 (left1-right1), dr1 (right1-left1), dl2 (left2-right2) and dr2 (right2-left2). The second block performs motion analysis on selected image frames (left1-left2) or (right1-right2) to generate point correspondences. The sub-modules in this module are arranged in a coarse-to-fine fashion (see Figure 4), so that less complex motion algorithms may be performed for initial fusion, and then refined gradually by performing higher-order motion analysis algorithms (we will later elaborate on our motion analysis algorithms to compute point correspondences). Consistency check is the crucial module of the system, and is explained pictorially in Figure 2. This module checks for the reliability of point correspondences generated by stereo and motion modules. Consider a point (x, y) in the left1 image, whose disparity is estimated as dl. Hence, location of the point in 3D can be given by (x, y, dl), treating disparity as the true height estimate. The 3D coordinates of point (x, y) on the right1 image will be $(x + dl, y, dr)$, where dr is the disparity calculated when right1 image is the reference. Motion analysis between left1 and left2 images produce point correspondences between these two images (this is performed by module 2). Hence, we will have an estimate of where the point (x, y, dl) moved in the next frame, denoted by $(x', y', d'l)$. $d'l$ represents the disparity at the point (x', y'), generated by stereo analysis between left2-right2 pair. This will also tell us the point's location on the right2 image, given by $(x' + d'l, y', d'r)$. $d'r$ is the disparity at the point $(x' + d'l, y')$, generated by stereo analysis of the pair right2-left2. Note that we have generated the motion correspondence of the point on right1 image, $(x + dl, y, dr)$ as $(x' + d'l, y', d'r)$ on the right2 image, without actually doing any motion analysis between

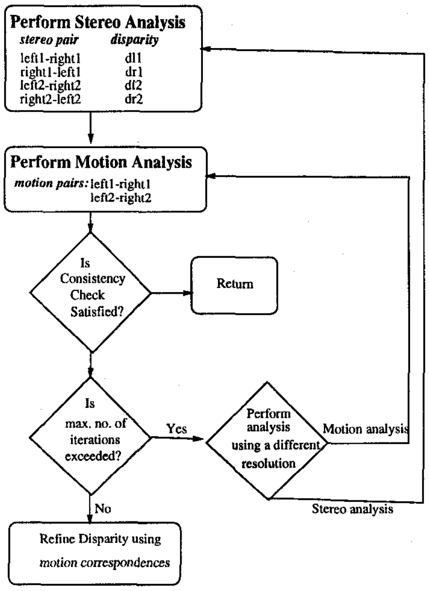

Figure 3: Algorithm

right1-right2. We know that the 3D vectors (assuming non-occluding points) connecting left1-left2 and right1-right2 are the same. One can calculate the dot product between the difference vectors, $(x' - x, y' - y, d'l - dl)$ and $(x' + d'l - x - dl, y' - y, d'r - dr)$, in order to evaluate the reliability of stereo correspondences (keeping in mind that initially, we assume the motion correspondence left1-right1 is reliable). Our aim is to iteratively refine the correspondence estimation (both motion and stereo) at different resolutions. Hence, if these vectors are far different, one can now use motion analysis on right1-right2 to get a correspondence for $(x + dl, y, dr)$ on the right2 image. The disparity on the left2 image is refined using this correspondence, and the process is repeated Nmax (or a threshold error chosen by user) times to see if the vectors reach an "agreement" (*stereo refinement*). If the threshold is not reached in the maximum no. of iterations, the system is started again by searching for a new motion match between left1-right1 (*motion refinement*). In each new cycle, the resolution level is increased for both stereo and motion analysis and a *finer* level is used. The description of the algorithm is also depicted in Figure 3. The system presented is interactive, allowing analysis at different multi-resolution levels. At anytime, user can chose any particular level of analysis in order to *refine* the results. We now present the multi-resolution stereo analysis and multi-resolution motion analysis algorithms used in our system.

2.1 Multi-resolution Stereo Analysis

Stereopsis, both in human vision and in remote sensing rely on the same principal of parallax or relative displacement. The primary difference being that in human stereopsis, the range of accurate depth information is on the order of several meters whereas for remote sensing applications depth information can be recovered over distances that are hundreds of kilometers due to the extremely long baselines available with satellite geometries [15]. The most difficult aspect of developing computational algorithms for stereopsis that match the intrinsic capabilities of the human vision system is the correspondence problem. i.e, locating the same feature in the sensor projected stereo datasets.

Correspondences are usually determined between a stereo pair using low-level features, regions, or multiple primitives [2, 7, 15, 12, 4, 13]. Correspondence information provides a disparity map which can be transformed into a depth or cloud-top height map using sensor geometry information. We follow a multi-resolution, hierarchical coarse-to-fine correlation-based approach in computing disparity between stereo pairs (also called reference and test image pairs) [13], implemented in parallel on a Maspar. In such an approach, high-level feature matching brings the images closer (initial fusion), so that the low-level matching is more efficient. The stereo analysis algorithm uses coarse-scale matching with large templates for estimating an initial fusion and then iteratively increases to fine-scale matching in analogy with psychological observations of human stereopsis (see Figure 4). In many multi-resolution stereo analysis approaches, the cost of generating images and matching over all levels even using an image pyramid is very high. In the Maspar parallel implementation, the image size remains constant and the matching template size is reduced. Matching is based on maximizing a normalized cross-correlation measure for a template centered around the pixel of interest in the reference image (starting template is generated by an automatic method). The search window in the test image incorporates epipolar and maximum disparity constraints and specifies the search region. The pixel in the test image corresponding to the maximum template correlation value is used for the cumulative disparity estimate at the current level. The updated dense disparity map is used to warp the test image so that searching at the next finer level can use a smaller sized template. The process is repeated until the changes in the disparity estimates are small or the finest level of the hierarchy is reached. In addition to this process, illegal disparities can be detected at any level of resolution and substituted with interpolated neighborhood disparities, along with disparity smoothing.

The main steps of the algorithm implemented for the Maspar are summarized as, *Preprocessing images*, *Automatic template size search*, *Initialize disparities*, *Warp test image*, *Determine image matches*, *Detect illegal disparities*, *Interpolate over outlier disparities* and *Smooth disparities*. The region-based measure used for stereo matching is a normalized mean and variance correlation, called match score given by,

$$\frac{\sum_{i,j}(x_{i,j} - \bar{x})(y_{i-k,j} - \bar{y}_k)}{\left[\sum_{i,j}(x_{i,j} - \bar{x})^2\right]^{\frac{1}{2}} \left[\sum_{i,j}(y_{i-k,j} - \bar{y}_k)^2\right]^{\frac{1}{2}}} \quad (1)$$

where $x_{i,j}$ and $y_{i,j}$ correspond to template pixels within the reference and test images respectively, $x_{i,j}$

is the grey level of the (i, j)th pixel within the template neighborhood and $y_{i-k,j}$ is the grey level of the (i, j)th pixel in kth search area of the test image. The values $\bar{x}$ and $\bar{y}_k$ are the corresponding mean values. For each pixel in the reference image, the match scores for all neighborhoods within the search area are computed. The pixel at the center of the search template with the highest correlation match score is selected as corresponding candidate pixel and the vector k gives an estimate of the disparity. The search window size and sub-pixel interpolation window size can be controlled. Vertical disparities are possible unless searching is constrained to (horizontal) epipolar lines, and the search direction can also be constrained. Unreliable match values can be filtered using thresholds based on local image variance. Please see [13] for more details.

2.2 Multi-resolution Motion Analysis

Motion perception in humans is a natural well-developed visual capability that is an essential survival skill. Over the last two decades there have been numerous developments in the area of motion analysis [8, 9, 1, 10, 13]. However, developing automatic computational motion analysis algorithms capable of handling various categories of motion has been a difficult task. Most of the work in motion analysis is based on the rigidity assumption where the shape of objects do not change over time. Motion of an object may involve complex changes of its structure itself by undergoing "non-rigid" motion. The rigidity assumption fails in numerous motion analysis situations, as many real-world objects are non-rigid. There are numerous areas where non-rigid motion is seen, including the atmospheric sciences, medicine, robotics or manufacturing. Typical examples of non-rigid motion behavior are also prevalent in nature including the dynamics of clouds and aerosols, dynamics of water waves and currents, dynamics of sand and soil, the motions of animals, plants and biological cells, or the deformation of flexible structures and industrial components. Fluid motion is perhaps the most complex that is commonly observed. There is usually no continuity constraint among neighboring particles since they move freely according to the underlying dynamics. Cloud motion is a special case of non-rigid motion, where there is partial fluid and partial solid motion, which is also termed "semi-fluid" motion behavior [13].

The objective of our system is to recover point correspondences between objects of time-varying imageries, using multi-resolution strategy. The *coarse-to-fine* motion analysis is naturally obtained by categorizing motion in a pyramidal structure of complexity. Rigid motion can be considered as the simplest of all, and fluid motion the most complex of all. In our approach, we consider different categories of motion to be in a multi-resolution (hierarchical, or coarse-to-fine) structure, as shown in Figure 4. One can perform motion analysis at any level of resolution. For example, rigid motion matching may be performed for initial fusion, and gradually raised to semi-fluid motion matching. Among the motion categories in the figure, elastic motion and fluid motion are the more general categories of non-rigid motion. Elastic motion deals with shape changes of continuous surfaces, and fluid motion deals with particle motion within fluids. Semi-fluid motion, on the other hand preserves the overall continuity of a surface during a fluid motion and allows movement of larger localized particle motions. Please see [10] for an elaborate description of the *simple-to-complex* motion categories.

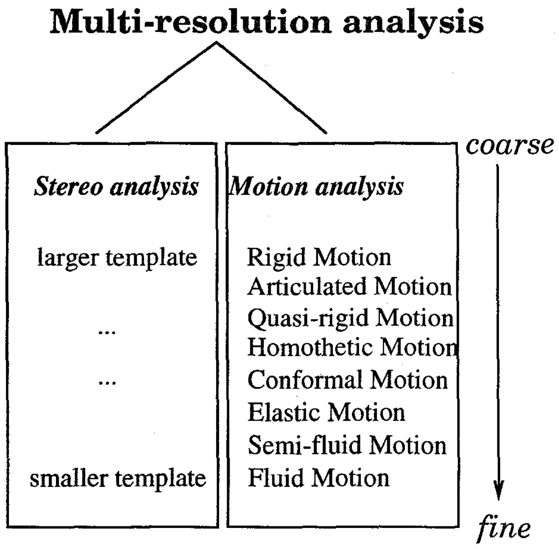

Figure 4: Multi-resolution Hierarchy

2.3 Experiments

Initial experiments have been performed on synthetic data so as to ensure working of the integrated system. Synthetic stereo image pairs were generated using NOAA/AVHRR data of Hurricane Andrew from August 25, 1992. Error between the estimated disparity and ground truth using mult-resolution stereo analysis alone was found to be very low with a mean of -0.274, a variance of 6.31, and a standard deviation of 2.51. We have also tested the algorithm by generating different stereo pairs with varied disparity ranges and have observed the mean error to be consistently closer to zero (pictorial results not shown due to space constraints). Synthetic motion analysis experiments were also performed on simulated time-varying data, and consistently low errors observed. Again, synthetic results are not shown due to space constraints. We now demonstrate the working of our integrated multi-resolution system on Hurricane Frederic dataset.

Stereo image pairs over several time steps, acquired by two geostationary satellites with synchronized scanning instruments are used in our analysis to generate cloud-top height estimation and cloud wind measurements. Four time sequential stereo pairs with a time interval of 7.5 minutes between each stereo pair and a lag of 15 seconds between the left and right pair were used. We ran this time-sequential stereo data on the integrated system. Automatic template selection algorithm is used to determine the starting template size (19X19 in this case) for the multi-resolution stereo analysis. The template size is gradually decreased

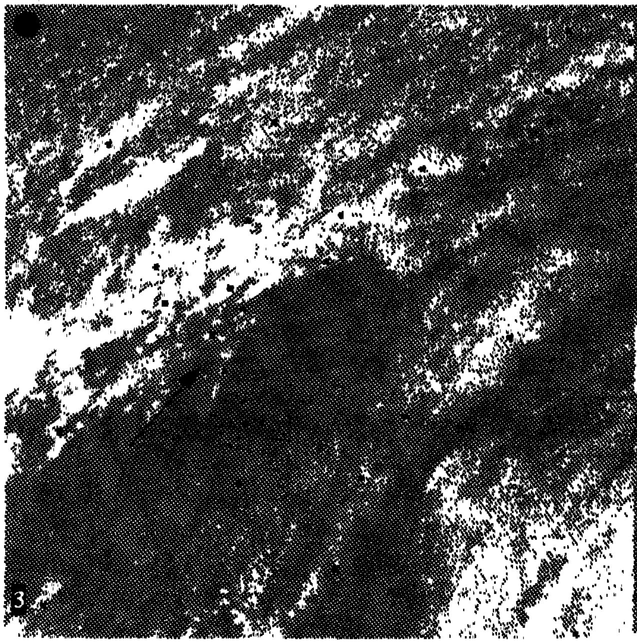

Figure 5: Intermediate result: Frame 3 of 4

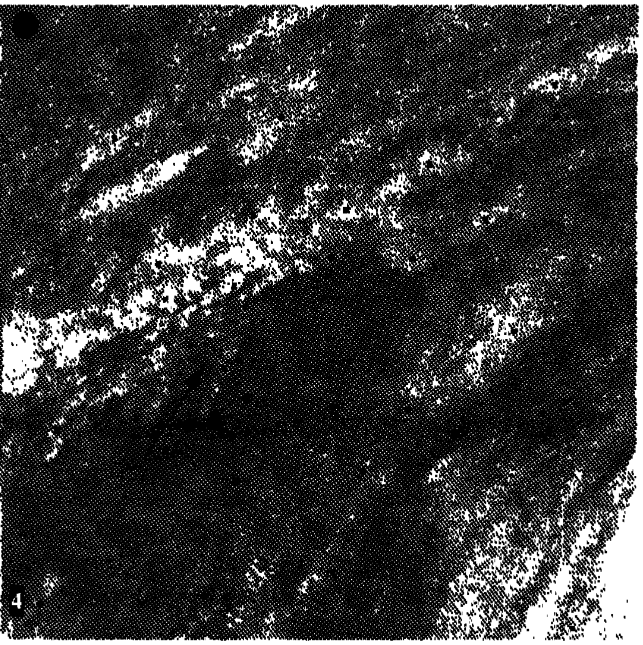

Figure 7: Intermediate result: Frame 4 of 4

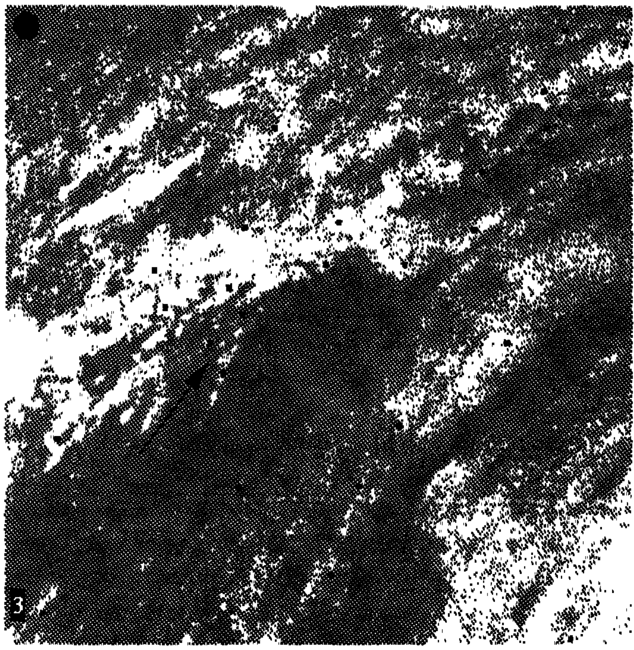

Figure 6: Final result: Frame 3 of 4

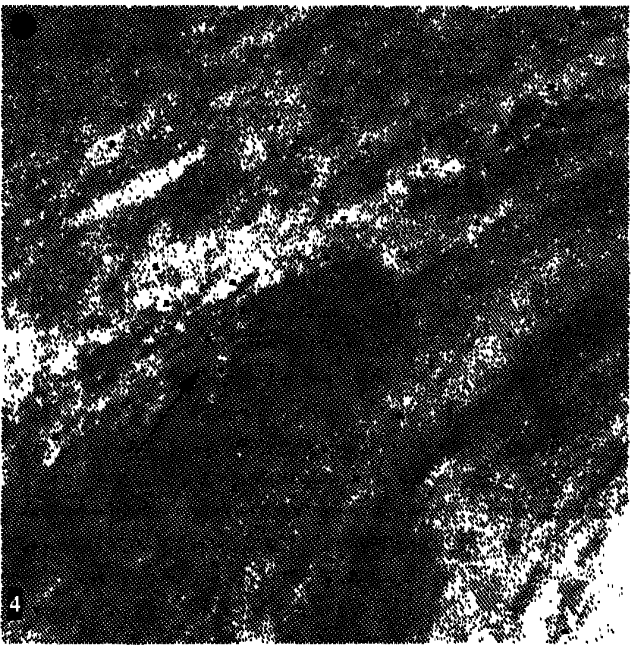

Figure 8: Final result: Frame 4 of 4

to (5X5), in coordination with the multi-resolution motion analysis (we followed the algorithm described in Figure 3). Similarly, motion analysis is started with rigid matching, and gradually refined to semi-fluid motion analysis (in coordination with the multi-resolution stereo analysis, following the algorithm). The results tally with the manual analysis of cloud heights and speed, performed by an expert in the field with a negligible error. There is less than 10% error in both height and speed computations, which can be attributed to human error in manual tracking. Figures 5 (frame 3 of 4 time-frames) and 7 (frame 4 of 4 time-frames) represent intermediate results, where the tracked points are shown by overlaying them on the left image (GOES-east). Figures 6 and 8 show the same frames at a finer resolution level. We can clearly observe better estimation of point correspondences, especially near wind tracers (check out point tracking near the red arrow), thus corroborating the integration of mult-resolution stereo and motion analysis.

3 Summary and Conclusions

An integrated system for multi-resolution stereo and motion analysis with sub-pixel accuracy is presented. The system is capable of handling varied categories of motion, arranged in a multi-resolution hierarchy. We have demonstrated the quality performance of the system on complex structures such as clouds. Extensions to the stereo analysis is proposed so as to include shape-from-X and other X-based techniques to add on to stereo analysis in producing better disparity at different resolution levels. Results of multi-resolution stereo analysis on synthetic data is excellent and can be used for interactive studies of cloud-top structures, owing to the parallel implementation for high-speed performance. Results of the multi-resolution motion analysis algorithm is also very promising and is comparable to the human analysis of cloud motion. Future work involves not only using illumination information but also other channels such as infrared as different modules of the system. New directions for incorporating robustness, motion segmentation, adaptive searching is necessary for enhancing the integrated system. Our future work includes a preprocessing phase for motion segmentation using expert knowledge. Model-based and physics-based analysis of stereo and motion is another direction to improve the existing integrated system. This work also involves development of new visualization tools and techniques for evaluating vision algorithms by enhancing the existing Interactive Image Spread Sheet (this has not been presented due to space constraints, however utilized as the 4th module in the integrated system).

References

[1] J. R. Bergen, P. Anandan, Keith J. Hanna, and Rajesh Hingorani. Hierarchical model-based motion estimation. *Proceedings of the European Conference on Computer Vision*, 1993.

[2] K.L. Boyer, D.M. Wuescher, and S. Sarkar. Dynamic edge warping: An experimental system for recovering disparity maps in weakly constrained systems. *IEEE Trans. Syst., Man, Cybernetics*, 21:143–158, 1991.

[3] T.M.H. Dijkstra, P.R. Snoeren, and C.C.A.M. Gielen. Extraction of 3D shape from optic flow: a geometric approach. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 135–140, June 1994.

[4] Lynne L. Grewe and Avinash C. Kak. Stereo vision. In Tzay Young, editor, *Handbook of PRIP: Computer vision*, volume II, pages 239–317. Academic Press, San Diego, California, 1994.

[5] A. F. Hasler. Stereoscopic measurements. In P. K. Rao, S. J. Holms, R. K. Anderson, J. Winston, and P. Lehr, editors, *Weather Satellites: Systems, Data and Environmental Applications, Section VII-3*, pages 231–239. Amer. Meteor. Soc., Boston, MA, 1990.

[6] A. F. Hasler, K. Palaniappan, M. Manyin, and J. Dodge. A high performance interactive image spreadsheet (IISS). *Computers in Physics*, 8(3):325–342, 1994.

[7] M. Holden, M. J. Zemerly, and J-P. Muller. Parallel stereo and motion estimation. In I. Pitas, editor, *Parallel Algorithms and Architecture for Digital Image Processing*, pages 1–57. John Wiley Sons Ltd, 1992.

[8] T. S. Huang. Motion analysis. In *Encyclopedia of Artificial Intelligence*, volume 1, pages 620–632. John Wiley and Sons, New York, 1986.

[9] C. Jerain and R. Jain. Polynomial methods for structure from motion. *Proc. 2nd ICCV*, Tarpon Springs, Florida, 1988.

[10] Chandra Kambhamettu, Dmitry B. Goldgof, Demetri Terzopoulos, and Thomas S. Huang. Nonrigid motion analysis. In Tzay Young, editor, *Handbook of PRIP: Computer vision*, volume II, pages 405–430. Academic Press, San Diego, California, 1994.

[11] Lingxiao Li and James H. Duncan. 3-d translational motion and structure from binocular image flows. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15, No. 7:657–667, July 1993.

[12] Suresh B. Marapane and Mohan Trivedi. Multi-primitive hierarchical (MPH) stereo analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16, no. 3:227–240, March 1994.

[13] K. Palaniappan, Chandra Kambhamettu, A. Frederick Hasler, and Dmitry B. Goldgof. Structure and semi-fluid motion analysis of stereoscopic satellite images for cloud tracking. *Proceedings of the International Conference on Computer Vision*, To be published.

[14] Sharath Pankanti, Anil K. Jain, and Mihran Tuceryan. On integration of vision modules. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 316–322, June 1994.

[15] H. K. Ramapriyan, J. P. Strong, Y. Hung, and C. W. Murray, Jr. Automated matching of pairs of SIR-B images for elevation mapping. *IEEE Trans. Geosciences and Remote Sensing*, 24(4):462–472, 1986.

[16] Ammon Shashua. Projective structure from uncalibrated images: Structure from motion and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, no. 8:778–790, August, 1994.

[17] J. Weng, N. Ahuja, and T.S. Huang. Optimal motion and structure estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, no. 9:864–884, September, 1993.