Chapter 24
Wide-Area Persistent Airborne Video: Architecture and Challenges

Kannappan Palaniappan, Raghuveer M. Rao, and Guna Seetharaman

Abstract The need for persistent video covering large geospatial areas using embedded camera networks and stand-off sensors has increased over the past decade. The availability of inexpensive, compact, light-weight, energy-efficient, high resolution optical sensors and associated digital image processing hardware has led to a new class of airborne surveillance platforms. Traditional tradeoffs posed between lens size and resolution, that is the numerical aperture of the system, can now be mitigated using an array of cameras mounted in a specific geometry. This fundamental advancement enables new imaging systems to cover very large fields of view at high resolution, albeit with spatially varying point spread functions. Airborne imaging systems capable of acquiring 88 megapixels per frame, over a wide field-of-view of 160 degrees or more at low frame rates of several hertz along with color sampling have been built using an optical array with up to eight cameras. These platforms fitted with accurate orientation sensors circle above an area of interest at constant altitude, adjusting steadily the orientation of the camera array fixed around a narrow area of interest, ideally locked to a point on the ground. The resulting image sequence maintains a persistent observation of an extended geographical area depending on the altitude of the platform and the configuration of the camera array. Suitably geo-registering and stabilizing these very large format videos provide a virtual nadir view of the region being monitored enabling a new class of urban scale activity analysis applications. The sensor geometry, processing challenges and scene interpretation complexities are highlighted.

K. Palaniappan (✉)
University of Missouri, Columbia, MO 65211, USA
e-mail: palaniappank@missouri.edu

R.M. Rao
Army Research Laboratory, Adelphi, MD 20783, USA
e-mail: raghuveer.rao@arl.army.mil

G. Seetharaman
Air Force Research Laboratory, Rome, NY 13441, USA
e-mail: Gunasekaran.Seetharaman@rl.af.mil

B. Bhanu et al. (eds.), *Distributed Video Sensor Networks*,
DOI 10.1007/978-0-85729-127-1_24, © Springer-Verlag London Limited 2011

Keywords Wide-area motion imagery · Wide field-of-view sensors · Very large format video · Persistent surveillance · Camera sensor arrays · High numerical aperture optics · Airborne imaging

1 Introduction

Wide-area persistent airborne video, also known as wide-area motion imagery (WAMI), wide-area persistent surveillance (WAPS), wide field-of-view (WFOV) imaging or very large format video is a newly evolving imaging capability that enables persistent coverage of geographical regions on the order of a few to tens of square miles. The enabling technology is the use of airborne camera arrays combined with computational photography techniques to integrate information from multiple cameras spatially, spectrally and across time in a consistent manner. Essentially a moving airborne camera array provides a denser sampling of the urban 4D light field or plenoptic function [1, 15]. The time-varying light field can be used in unique ways for large scale detailed 3D scene reconstruction, monitoring activity patterns of vehicles, people and animals, rapid change detection or providing continuous situation awareness for remote operations at high resolution. A network of such airborne camera arrays would be ideally suited for exploring a range of novel applications, previously considered technically infeasible or cost prohibitive, in urban monitoring, planning and design, ecological surveys, agriculture, traffic analysis, law enforcement, critical infrastructure protection, event security, emergency response after natural disasters (i.e., floods, hurricanes, tornadoes, forest fires, landslides, earthquakes, tsunamis), monitoring environmental disasters from anthropogenic activities (i.e., oil spills, pollution, mining, deforestation), search and rescue, border patrol, tele-operation, and defense.

Persistent wide-area airborne imaging typically uses a continuous circular flight path in a fixed 3D plane perpendicular to the local ground plane. Figure 1(a) shows an example persistent flight path along with the ground projected trajectory of an elevated point on a building in the scene. The varying viewpoints of the same stationary (or nearly stationary) object induces an apparent motion or *wobble* of those objects that are above the ground plane, with taller objects having a larger wobble. The parallax induced wobble poses both human factors and computational challenges for the visual interpretation of wide-area persistent surveillance slow video. An eight-camera optical array constructed by Persistent Surveillance Systems that was used to collect the aerial imagery described in this paper is shown in Fig. 1(b). The configuration of the cameras, focal lengths, pointing directions and overlap regions between adjacent camera FOVs are specified in Fig. 1(c) and (d). An airborne camera array can be used in persistent or stare-mode as well as survey or along-track mode. The latter mode is useful for rapidly sampling a large geographical region and can operate at approximately 600 square miles per hour using the eight-camera array shown (4 mile wide swath and speed of 150 mph). For the same swath width and airborne platform speed, a short video sequence or *cliplet* provides coverage of a given location for nearly 100 frames in survey mode. The visible channel imagery

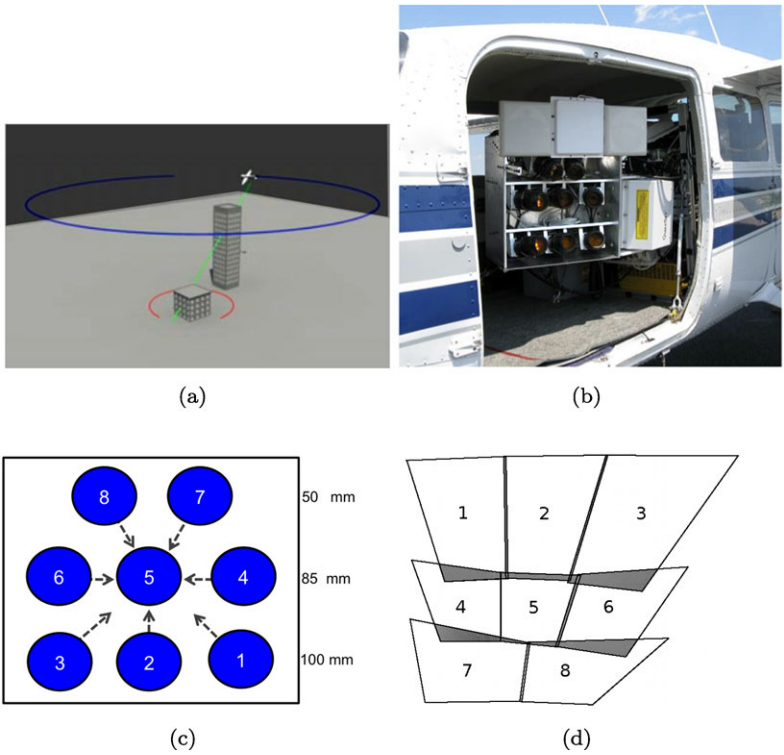

Fig. 1 Persistent wide-area airborne data collection using an eight-camera array. (**a**) A wide area circular orbit (*in blue*) of the aircraft and ground-plane trajectory (*in red*) of a 3D corner point on the lower building with geometric occlusions shown. (**b**) Eight-camera imaging array with varying optics built by Persistent Surveillance Systems shown mounted inside a long-endurance Cessna C-207 aircraft (photo by Ross McNutt). (**c**) Camera numbering, focal lengths and pointing directions for physical layout of the camera array. (**d**) Image-plane numbering of the projected camera views showing inter-camera overlapping regions where seams are likely to occur in the geo-registered image

can be augmented with other sources of information including infrared imagery for nighttime coverage, and hyperspectral imagery to characterize material properties for object identification. The importance of synthetic aperture radar (SAR) and moving target indicator (MTI) radar for synergistic all-weather, day-and-night coverage in wide-area surveillance was recognized early-on by [9]. In this paper we focus on the persistent mode of observation using visible channel imagery.

Each camera in the eight-camera array produces an 11 megapixel 8-bit visible channel grayscale image at one to four frames per second and of size 4096×2672 that is geo-registered to an $8K \times 8K$ or $16K \times 16K$ image mosaic. At the higher spatial resolution and higher temporal sampling this data volume is about four terabytes per hour or the equivalent of about 120 UAV standard definition 30 frames per second video streams. At a platform altitude of 1370 m (about 4500 ft) the nominal ground resolution, ground sampling or separation distance (GSD) is between 20 to

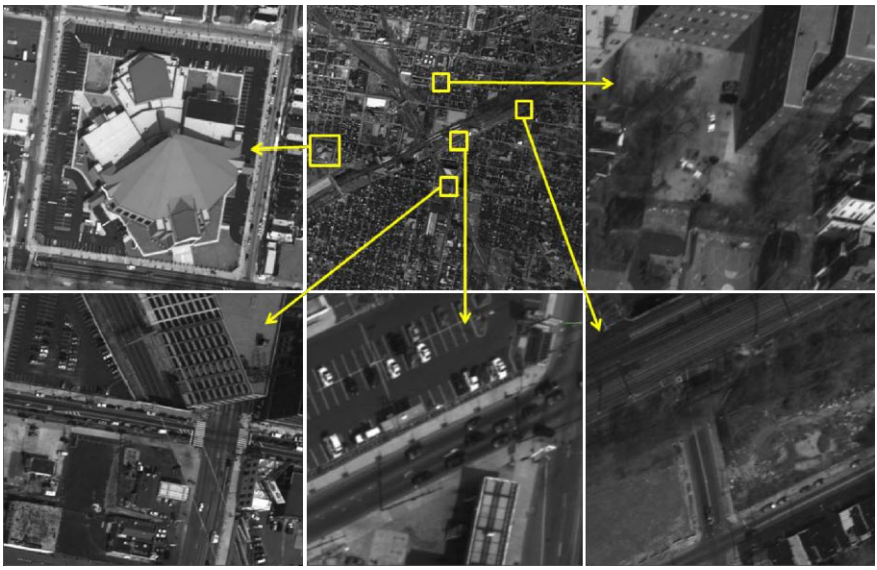

Fig. 2 Wide area image of North Philadelphia with area of interest inserts showing the high resolution available in the geo-registered mosaic. Image provided by Ross McNutt

25 cm (0.66 to 0.82 ft) per pixel; at an altitude of 2440 m (about 8000 ft) the GSD reduces to about 50 cm (1.64 ft) in the central part of the image with decreasing resolution towards the image periphery. At the lower altitude each mosaicked image frame covers about four square miles and at the higher altitude about 16 square miles. Examples of images collected at both resolutions, for several geographical regions and up to two frames per second are used in this paper. Figure 2 shows a portion of a sample 16K×16K wide area image over North Philadelphia, Pennsylvania taken on March 13, 2008 at the lower altitude. We will refer to this wide-area persistent video data set as the Philadelphia sequence. The area of interest inserts in Fig. 2 show zoomed views and the high resolution available in the wide-area image mosaic. Visualizing and analyzing such large time-varying data sets in an interactive manner requires careful software design of data structures, display tools and human computer interfaces to improve usability, data access and information presentation [2, 8, 12, 21].

The large oblique or wide field-of-view (WFOV), circular orbit of the airborne platform, unsteady ground plane, time-varying occlusions and stabilization for parallax mitigation leads to a number of challenges in developing robust visual feature-based object tracking algorithms [25]. The fact that a large area is being continuously sampled using a camera array leads to a new paradigm for analyzing such videos, largely stemming from non-uniformity in instantaneous optical characteristics and platform motion. At any given instant, the areas in the center of the scene, usually around the fixation ground-point, will be observed at the highest resolution, while the image computed across the rest of the WFOV will originate from resolution-limited highly oblique line-of-sight data using the camera-array configu-

ration in Fig. 1. Another significant challenge is the scale of the activity occurring in the scene, which is similar in complexity to simultaneously analyzing a dense distributed network of hundreds to thousands of airborne or ground-based video cameras. Wide-area motion imagery of urban areas produce tens of thousands of interrelated spatio-temporal events particularly in relation to moving objects that are interacting in a highly non-linear dynamical fashion. Multiobject identification, automatic tracking including detection, track initiation, track management, mitigation of distractors and track termination applied to such a large collection of moving objects poses numerous computational, algorithmic and database challenges.

The spatially varying optical transfer function across the WFOV, non-uniform spatial resolution, low frame rate, and high parallax of urban structures using airborne camera arrays often do not satisfy the usual assumptions made by many existing vision algorithms. Well known approaches for image registration, video stabilization, optical flow analysis, and structure from motion algorithms [5, 22, 23, 26, 32–35, 38] have to be revisited because of the spatially varying optics, inherent heterogeneity of the large areas being monitored combined with the low temporal sampling frame rate and geometric complexity of the scene. Some applications of (non-persistent) WFOV images collected using bursty sampling to address the low frame rate limitation for vehicle tracking and traffic pattern analysis using aerial imagery are described in [13, 28]. A unique way of exploiting the short time difference between dual sensor (panchromatic and color) images acquired by satellites such as QuickBird to estimate MTIs for wide-area periodic surveillance is described in [7].

In this chapter we focus on persistent WFOV image sequences and the challenges associated with their analysis which are broadly described using examples. Figures 3 and 4 show perspective changes in building shape and occlusion events that make frame-to-frame registration, stabilization, recovering 3D structure and moving object tracking tasks more challenging. Both figures are from the Philadelphia wide-area motion imagery sequence. In Fig. 3 the view of the church building changes from an oblique view to a more nadir view across the four frames (700×500 pixels) that are each 10 sec apart starting at frame 45709. In Fig. 4 the view of the triangular office building and the geometric occlusion of surrounding structures is seen across these selected four frames (1800×1600 pixels) that are sampled 40 sec apart also starting at frame 45709. Tracking objects through such long occlusions may be feasible by combining additional information from ground-based video networks, or using multiple wide-area and other complementary airborne platforms to improve total coverage. Feature extraction, texture descriptors and point correspondence methods [11, 17, 27, 34] can be easily overwhelmed by the large number of spurious matching points with similar configurations detected at different geo-altitudes and the symmetric, repetitive structure of buildings as apparent in Fig. 4.

Objects moving steadily across the WFOV can persist and stay visible for long durations with intermittent to extended occlusions. However, stabilization, mosaicking, moving object detection, blob segmentation, track initiation, reacquisition, occlusion handling and pair-wise relations between moving targets are complex vision tasks even for regular airborne or ground-based video [3, 4, 10, 24, 29, 36, 37] that

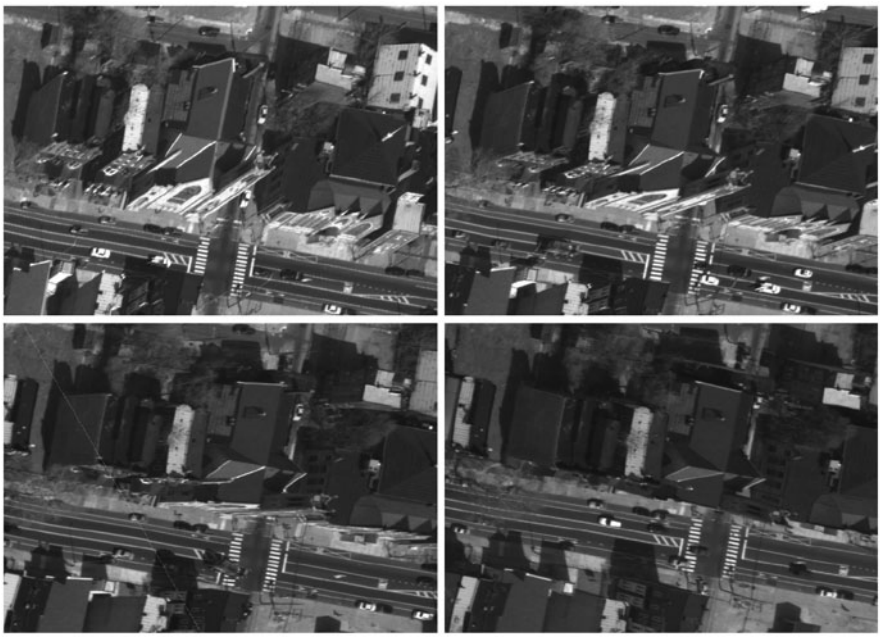

Fig. 3 Viewpoint induced changes in the appearance and pose of building structures across short time periods can be seen in these four regions from the Philadelphia sequence (10 sec apart)

need to be further extended to the WAMI domain to support exploitation of city-wide and region-wide scene activity analysis. We use radiometric and geometric characterization, tracking and pose-stabilization to illustrate some of the challenges in the exploitation of WAMI. A more detailed description of using feature fusion to improve vehicle tracking in WAMI is described in [25]. It is not our intent here to focus on any single area each of which merits a dedicated review. Instead, our aim is to introduce the overall architecture of wide area imaging with steadily moving camera arrays and describe some of the newly enabled opportunities along with associated challenges.

2 Spatio-temporal Reflectance Variations

Wide-area motion imagery system performance depends on a large number of factors that affect image characteristics. The following list is a collection of the more important factors including: the number of cameras in the array, their relative poses with respect to each other, the lens optics and FOVs, multiple camera calibration, radiometric balancing, geo-registration, mosaicking, frame rates, target size, target speed, clutter, weather, sun-angle, the number of channels and modalities (color, IR, etc.), the platform altitude, GPS accuracy, inertial measurement unit (IMU) accuracy for measuring pose of the aircraft platform, on-board data storage, down-

Fig. 4 The tall triangular building creates geometric occlusions of the road, ground-plane level structures, vehicles and pedestrians due to viewpoint induced *wobble* that is proportional to building height

link bandwidth, attributes of the ground processing system, and so on. Radiometric balancing and geometric registration across the camera array are essential to produce high quality WAMI mosaics. Geometric registration and radiometric balancing problems give rise to seams in the overlap regions of the multi-camera image mosaic as illustrated in Figs. 5 and 6 respectively. Such deficiencies in image quality adversely affect downstream image processing and scene analysis modules including feature extraction, feature tracking, depth reconstruction and object identification. Note that the shape of the geometric seams or overlaps between adjacent image planes changes with time based on the platform position and pose, and intrinsic and extrinsic camera-array configuration (see Fig. 1). We use a recently enhanced version of the Kolam software tool for interactive visualization of the very large WAMI mosaics [21].

The combination of sensor behavior and scene changes across short time intervals can be characterized by looking at the spatio-temporal variation of measurements such as geo-registration and radiometric accuracy between cameras. Registration and stabilization problems, significant parallax variations, changes in view-

(a) (b)

Fig. 5 The geometric seams across adjacent camera planes and the induced image distortions in vehicle and building structures can been seen in WAMI sections of (**a**) Philadelphia, PA, March 2008, and (**b**) Cedar Rapids, Iowa, June 2008

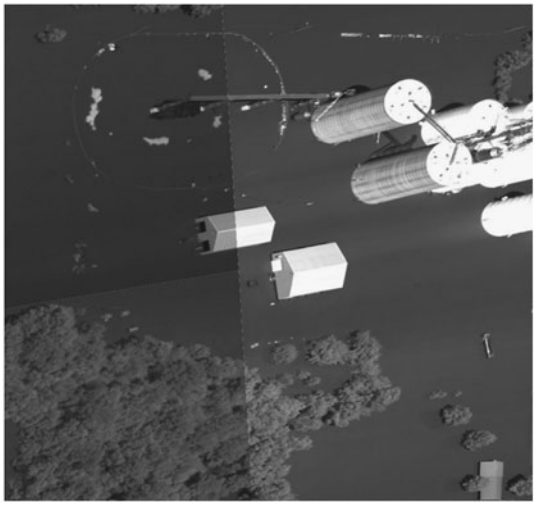

Fig. 6 The radiometric seams between three adjacent camera image planes requiring correction are evident in this image showing flooding around a farm in Oakville, Iowa, June 2008

point and changes in illumination are a few of the factors that can cause abrupt changes in scene quality over time. We use the Wasserstein distance (equivalent to the Earth Mover's Distance (EMD) under certain conditions) to measure the spatio-temporal variability in the intensity histogram distributions. Interframe differences can be measured in terms of pixel-level gradient changes, regional-level histogram changes, optical flow motion vectors, frame-level feature statistics, or other appropriate video-based measures. Distribution based techniques provide global infor-

mation about an image and are less sensitive to small camera motions and object motions compared to spatial interframe differences. Let $P_a(y)$ and $P_b(y)$ be two normalized density functions (i.e., histograms) and $F_a(y)$ and $F_b(y)$ be their corresponding cumulative distributions. Then the linear Wasserstein distance, W_1, between $P_a(y)$ and $P_b(y)$ across an intensity range G is defined as [20],

$$W_1(P_a, P_b) = \int_0^G \left| F_a(y) - F_b(y) \right| dy. \tag{1}$$

Since the maximum value of the difference in the cumulative distributions is one, the maximum value of the integral is G which can be used as a normalization factor. In the discrete approximation to the Wasserstein distance using summations instead of integrals, we need to take into account the histogram or density function bin size in the normalization factor. If the histogram is sampled using a bin size of $\triangle h$ then W_1 should be normalized by $(G/\triangle h - 1)$.

Figure 7 shows temporal changes in the scene reflectance function over a short time period of 105 seconds for three different geospatial regions from wide-area persistent imagery of Juarez, Mexico on August 26, 2009, which we refer to as the Juarez data set. In this case we selected three 512×512 regions or image blocks centered at pixel locations [5888, 5888], [6912, 6912], [7936, 7936] across 105 frames starting from frame 48051. The first row shows a representative image sampled from the center of each corresponding x–y–t spatio-temporal block of image data. The second row shows the temporal variation of the x–y block graylevel intensity histograms as 3D surface plots. The third row shows the variability of a horizontal line profile (at row 256) in the region as a spatio-temporal x–t slice with time in the vertical dimension. If the images are perfectly registered and compensated for viewpoint changes then we would expect vertical lines instead of sinusoidal patterns. The fourth row shows the temporal variation in the average x–y block graylevel intensity for each region. The vertical axes cover slightly different ranges but the rapid change in mean reflectance intensity for each spatial region is readily evident and is primarily due to viewpoint changes resulting in the appearance change of objects. In the ideal case we would expect a horizontal curve if the mean reflectance for the block remained constant.

The fifth row shows intensity histogram differences measured using the Wasserstein distance with the normalization factor included. The red curve plots the histogram difference between each pair of consecutive frames whereas the blue curve plots the histogram differences with respect to a reference reflectance histogram based which in this case is based on the center frame, with a small blue circle on the x-axis marking the reference frame number 53. The discrete approximation to the Wasserstein distance is normalized by the number of bins in the histogram and measures the change in the illumination distribution across the scene. The vertical axes have different ranges to highlight the variability in the reflectance distribution between images separated by short time periods. When we look at adjacent frames the variability in illumination as measured using the Wasserstein distance is low, however, the illumination differences between frames separated by even a few seconds can be quite large as shown by the blue curves in row 5. In this case there were no

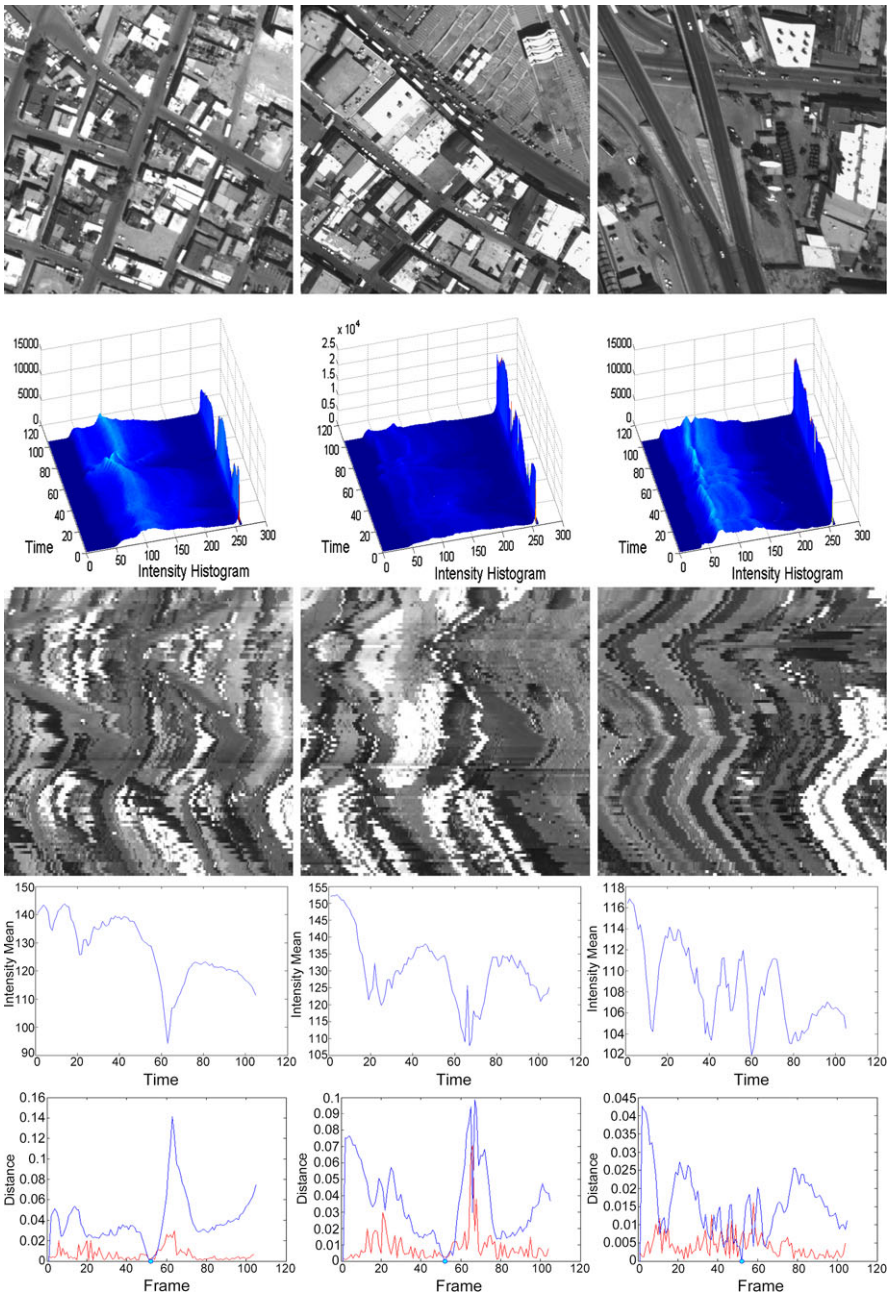

Fig. 7 Reflectance change in the Juarez data set for three 512 × 512 × 105 image blocks. Row (**a**) ROI from center frame; (**b**) 3D plot of x–y block intensity histograms over time; (**c**) spatio-temporal x–t slice for row 256; (**d**) x–y block mean intensity over time; (**e**) Wasserstein distance plot between intensity histograms for consecutive frames (*red curve*) and with respect to the middle reference frame (*blue curve*)

clouds in the scene so the major contribution to the changes in the reflectance distribution are more from viewpoint change than misregistration errors. All three cases show some gradual cyclical changes over the 105 frames or 105 seconds of video, apparent both in spatio-temporal slice drift (row 3) and reflectance variation plots (rows 4 and 5), especially using the Wasserstein distance measured with respect to the center image frame. The rapid spatio-temporal variation in reflectance highlights the fact that in wide-area motion imagery the appearance of objects change significantly with viewpoint and that such changes need to be modeled for accurate extraction of visual scene information for object tracking and object structure.

3 Wide Aperture Imaging Model of Camera Arrays

It is well established in physical optics that imaging through a lens is governed by the Rayleigh principle and various factors associated with geometrical optics and the design of lenses. The optical energy originating from a distant object captured by the lens is approximated as a paraxial parallel beam. The total energy captured in such an imaging system is proportional to the cross-sectional area of the *aperture*. The image of any paraxial parallel beam is approximated by a blurred spot whose diameter is inversely proportional to that of the aperture. The relationship is expressed in the form

$$d_B = 2.4\lambda(F/\#) = 2.4\frac{\lambda f}{D_a}, \tag{2}$$

where f is the focal length, D_a is the aperture diameter, d_B is the blur-spot diameter, and λ is the wavelength of light captured by the lens. The aperture is generally described in photography as $F/\#$ (the F-stop number) and is proportional to the inverse of the numerical aperture or resolving power of a lens. It is desirable to have a large aperture (i.e., low $F/\#$) to produce optical images with higher effective ground sampling distance or resolution. The mutually interdependent constraints require tradeoffs between large physical apertures, long focal lengths and limited field-of-views (FOV), that is telephoto lenses, when imaging distant objects. Larger aperture lenses with shorter focal lengths, which would be preferred, are difficult to design; such lenses are usually expensive and likely to introduce severe image distortions. Additional factors that govern the choice of lenses require a combined analysis of pixel size of the imaging light sensor, desired object-spot diameter (OSD), stand-off distance between the object and camera, and the required FOV to cover the scene of interest. Accommodating depth of field, or variations in the extent and distance of the object from the camera, is also an important factor in selecting appropriate lenses.

The design of more flexible imaging systems using camera arrays, lens arrays, coded apertures, catadioptrics combining lenses and mirrors, spatial light modulation and a variety of other techniques using physical devices and digital processing is an active area of research. A variety of such computational camera arrays and optical systems have been built [14, 15, 18, 19]. Very large format or wide area

imaging systems are based on the design of a wide field-of-view (WFOV) imaging system using an appropriately arranged array of smaller FOV cameras to achieve the highest optical resolution across the greatest FOV with the best possible temporal sampling. We illustrate the key ideas geometrically using a 2D cut through the principal cross-section of a camera array, showing how the net angle of view (AOV) can be increased using a specific spatial arrangement of cameras. In the case shown in Fig. 8 they have been placed on the circumference of a circle. This architecture can be readily extended to the 3D case for building the corresponding real-world camera-array system. The goal is to construct a distributed aperture imaging system composed of a camera array with multiple focal planes that produces a view equivalent to that acquired by an ideal WFOV camera with a very large focal plane detector array.

3.1 Seamless Stitchable Camera Arrays

Desirable criteria and considerations for designing a WFOV multi-camera array are listed here as a set of guidelines. The design should ensure that the net WFOV is a simply connected set (without holes) in a suitably represented manifold and without discontinuities in pan, tilt and azimuth angles in the final image. Overlaps between individual camera FOVs should be minimal, for the obvious reason of maximizing the collective WFOV extent. Multi-camera-array calibration should be of similar complexity to calibrating a single-camera imaging system. The rigid-and-static relative position and orientation between the cameras can be set to any desired degree of precision and calibrated. Without loss of generality we assume a minimum separation distance between objects in the scene and the camera array. Image registration, if required for producing the single-perspective image, should be simple, easy to compute, and be applicable across a wide range of scene conditions. That is, any explicit assumptions on the nature of the 3D scene should be minimal for registering across the camera-array views to produce the equivalent WFOV seamless mosaic image. In addition, we want to minimize the total computations needed to reconstruct the WFOV stitched image and reach system performance and accuracy specifications.

One basic realization of such a design is shown in Fig. 8 using three identical cameras C, D, E, each with a 24° FOV, equally separated on a circular arc such that the three-camera array as a whole produces a net WFOV of 72°. Camera D has been placed such that $AD \parallel BC$ and E has been placed such that $BE \parallel AC$. Collectively, the FOVs $\measuredangle ADB + \measuredangle ACB \mapsto \measuredangle AdB$, and $\measuredangle ACB + \measuredangle AEB \mapsto \measuredangle AeB$; from which it can be shown that the net WFOV at the effective imaging array focal plane, c is, $\measuredangle ADB + \measuredangle ACB + \measuredangle AEB \mapsto \measuredangle AcB$. In principle, it is possible to transform the image captured by the cameras, E, C and D to their equivalent counterparts as seen from location c. Such a transformation can be separated into two distinct cases: (1) distant objects, and (2) nearby objects with respect to the camera array.

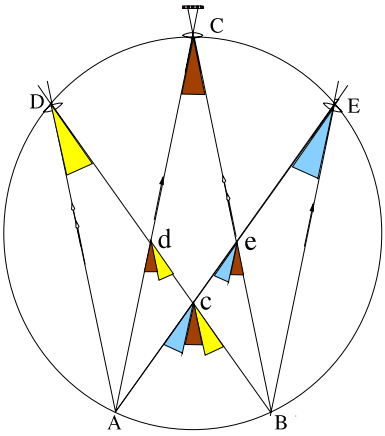

Fig. 8 Three identical cameras, C, D and E are placed such that, lines $AD \parallel BC$, and $BE \parallel AC$. Line segment AB is considered a design parameter governed by other considerations

Let the size of each pixel be $\triangle_x \approx d_B$, where d_B is the blur-spot diameter. Let the distance between the object and the pupil of the camera through which it is seen be q. Then $d_O = q\triangle_x/f$ defines the size of an object patch seen by any pixel. If the net displacements $\|Ec\|$, $\|Cc\|$, and $\|Dc\|$ are small ($\ll d_O$), then an acceptable approximation is that the images recorded at c are identical to those seen at E. Thus, it is acceptable to trivially inherit the image from camera E with minimal postprocessing, in order to compute its contribution to the WFOV (mosaic) image that would be recorded at c. A similar reasoning applies to the images acquired by cameras C and D.

3.2 Geometric Properties of WFOV Imaging Arrays

We describe the global non-linear nature of perspective imaging and local linear approximations suitable for the analysis of WFOV multi-camera-array images following the notation for single-camera modeling [6, 30, 31]. In general, video cameras project a certain object point $\mathbf{X}$ located on opaque objects onto an image point $\mathbf{x} = (x, y, z = f)$ in the image plane. The image plane is uniquely characterized by the focal length f of the camera, C, expressed by the equality $Z = f_C$. The projection model of the image sensor is either *perspective* or *orthographic* depending on the lens characteristics and physical dimensions of the image sensor in comparison with the focal length of the lens and distance to the object. Loss of depth information is inevitable in both types of projections. The intrinsic geometric models for an intensity camera are illustrated in Figs. 9 and 10. A WFOV imaging system, as a whole, mimics a perspective imaging system. However, it produces images that are locally orthographic since the individual cameras are built with telephoto lenses. We refer to the telephoto lens-based images as *weakly perspective*, or *piece-wise orthographic*. Such an insight can be fully exploited in a framework similar to small-signal analysis used in modeling circuits built with non-linear elec-

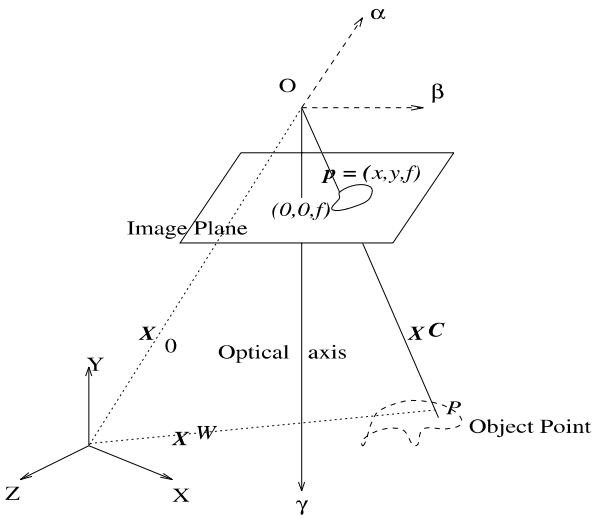

Fig. 9 A single-camera perspective imaging system of an object point P and its projection onto the imaging plane

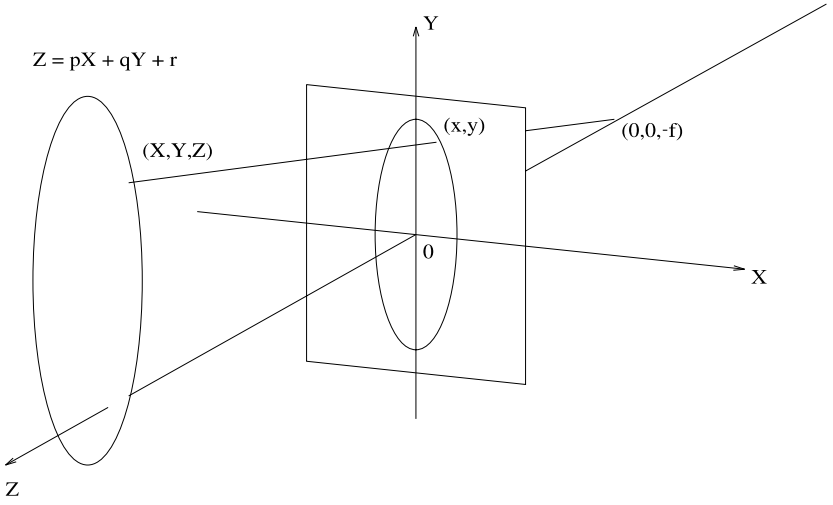

Fig. 10 A simple orthographic imaging system model

tronic devices. Temporal analysis of small patches can therefore be studied as if they are related by affine transformations. At the same time, global analysis of the WFOV image will almost always exhibit large departures from an aggregate linear model.

A clear insight into the underlying geometric properties will help approximate the complex WFOV analysis using suitably partitioned fields of view, each modeled

as an affine mapping. The irreversible loss of depth information introduced by the underlying perspective projection can be expressed as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{P} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \quad \text{where } \mathbf{P} = \begin{bmatrix} \frac{1}{\lambda} & 0 & 0 \\ 0 & \frac{1}{\lambda} & 0 \\ 0 & 0 & \frac{1}{\lambda} \end{bmatrix}, \ \lambda = \frac{Z}{f_C}, \tag{3}$$

with $z = f_C$ and $\lambda \gg 1$. Any point on the ray defined by the vector $\mathbf{X} = [X\ Y\ Z]^T$ which is $a\mathbf{X}$, $\forall a, a \neq 0$, projects to the same image point $\mathbf{x} = [x\ y\ f]^T$. The projection is non-invertible; thus, given $\mathbf{X}$ one can determine $\mathbf{x}$ but not the opposite. However, given a point $\mathbf{x}$ on the intensity image, $\mathbf{X}$ is constrained to a line (of points) passing through the focal point $\mathbf{O}$ and the image point $\mathbf{x}$. In order to relate measurements between multiple cameras we use superscripts to describe the frame of reference and subscripts to identify the object of interest. The notation $\mathbf{O}_C^W$ is used to describe the position of $\mathbf{O}_C$ measured with respect to the world coordinate system W. Given the absolute position $\mathbf{X}^W$ of a point, X, measured with respect to the world coordinate system, both $\mathbf{X}^C$ and $\mathbf{X}^W$ are related as follows,

$$\left[\begin{array}{c} X^W \\ Y^W \\ \hline Z^W \\ 1 \end{array}\right] = \mathbf{T}_C^W \left[\begin{array}{c} X^C \\ Y^C \\ \hline Z^C \\ 1 \end{array}\right], \tag{4}$$

with the six *extrinsic camera parameters* collected together in the matrix,

$$\mathbf{T}_C^W = \left[\begin{array}{ccc|c} \boldsymbol{\alpha}_C^W & \boldsymbol{\beta}_C^W & \boldsymbol{\gamma}_C^W & \mathbf{O}_C^W \\ \hline 0 & 0 & 0 & 1 \end{array}\right], \tag{5}$$

where $\boldsymbol{\alpha}_C^W, \boldsymbol{\beta}_C^W$ and $\boldsymbol{\gamma}_C^W$ are the direction cosines of the X, Y and Z axes of the camera, and $\mathbf{O}_C^W$ is the origin of the camera coordinate system. The matrix $\mathbf{T}_C^W$, is uniquely characterized by these extrinsic camera parameters and is always invertible. These parameters are easily calculated when the position and orientation of the camera is known with respect to the absolute coordinate system. They can also be extracted using calibration techniques. From (3) and (5), it clearly follows that given $\mathbf{x}^C$, additional information is required to uniquely locate $\mathbf{X}^W$ along the projective ray,

$$\left[\begin{array}{c} X^W \\ Y^W \\ \hline Z^W \\ 1 \end{array}\right] = \mathbf{T}_C^W \left[\begin{array}{c} \lambda x^C \\ \lambda y^C \\ \hline \lambda f_C \\ 1 \end{array}\right] \equiv \mathbf{T}_C^W \left[\begin{array}{c} \lambda\, \mathbf{x}^C \\ \hline 1 \end{array}\right]. \tag{6}$$

In certain circumstances, it is desirable to model the camera as an orthographic projection. Practical cameras are inherently perspective. However, favorable conditions occur when the focal length of the camera is much larger than the diagonal size of the video sensor and/or the lens' diameter, or the distance to the object is very large. The non-linearity due to perspective imaging is uniformly distributed since

the lens covers only a very narrow beam of light, consisting essentially of almost parallel lines. The orthographic projection approximation is illustrated in Fig. 10. In contrast with the previous projective camera model, here we have moved the perspective projection point to $(0, 0, -f)$ without any loss of generality. In addition we represent a generic object point conveniently as the sum of an object centric coordinate system and a suitably defined translation. The object centered coordinate system is defined such that its axes are parallel to the respective camera coordinate system and its origin is located at $\widehat{Z}$, which is the average depth or distance to all object points (i.e., z-component of the object centroid). The basic equations are similar to that of the perspective imaging model but differences can be better emphasized if object points are described with respect to an object centered coordinate system.

Let a point in the scene be represented in the object centered coordinate system as $\mathbf{X}^C = (X^O, Y^O, Z^O) + (0, 0, \widehat{Z}^C)$, where $\widehat{Z}^C$ is the distance from the camera to the object centroid, then

$$x = \frac{fX^O}{f + Z^O + \widehat{Z}^C} = \frac{fX^O}{(f + \widehat{Z}^C)} \left(1 + \frac{Z^O}{\widehat{Z}^C + f}\right)^{-1}, \tag{7}$$

$$y = \frac{fY^O}{f + Z^O + \widehat{Z}^C} = \frac{fY^O}{(f + \widehat{Z}^C)} \left(1 + \frac{Z^O}{\widehat{Z}^C + f}\right)^{-1}. \tag{8}$$

In particular, when either $\widehat{Z}^C$ (the average depth) or f takes on very large values, the resulting projection is of the form

$$x = \lim_{\substack{\widehat{Z}^C \to \infty, \\ f \to \infty}} X^O \left(1 + \frac{\widehat{Z}^C}{f}\right)^{-1} \left(1 + \frac{Z^O}{\widehat{Z}^C + f}\right)^{-1} \approx X^O \left(1 + \frac{\widehat{Z}^C}{f}\right)^{-1}, \tag{9}$$

$$y = \lim_{\substack{\widehat{Z}^C \to \infty, \\ f \to \infty}} Y^O \left(1 + \frac{\widehat{Z}^C}{f}\right)^{-1} \left(1 + \frac{Z^O}{\widehat{Z}^C + f}\right)^{-1} \approx Y^O \left(1 + \frac{\widehat{Z}^C}{f}\right)^{-1}. \tag{10}$$

Both X^O and Y^O components of the position vector $\mathbf{X}^O$ are scaled by the same amount, and the scale is independent of the exact Z position, $Z^C = Z^O + \widehat{Z}^C$, of the object points. In essence, this is a scaled orthographic projection that preserves various second order geometrical properties, and it is in fact affine in nature:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \left(1 + \frac{\widehat{Z}^C}{f}\right)^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X^O \\ Y^O \\ Z^O + \widehat{Z}^C \end{bmatrix}. \tag{11}$$

Note, we assumed that the origin of the object centered coordinate system was located along the camera optical axis and that Y^O and X^O are within the viewing range.

3.3 Physical Considerations Governing Camera-Array-based WFOV Virtual Focal Planes

In general the camera array is constructed using a rigid mechanical structure in which the cameras, a GPS and a high-precision orientation sensor (inertial measurement unit) are mounted together. The calibration of the pair-wise relationships between their pointing angles is commonly referred to as boresight-offset calibration. Since the cameras D and E shown in Fig. 8 have already been rotated, the rays incident on their pixel plane will suffer minimal loss due to the expected oblique incidence experienced by a camera pointing in the original Z direction of the camera array. There are two approaches for using the images obtained by cameras D and E. The first one is to associate with each pixel in its own image a unique direction cosine characterizing the underlying line of sight, and simply map that pixel value to a pixel location on the virtual-image place constructed by extending the image plane of C. Such a process will imply non-uniform sampling across the image plane. The second approach is to start off with a uniformly sampled grid of the virtual-image plane, characterize each pixel by a line of sight, then fetch supporting measurements from the images associated with cameras C, D and E and interpolate the values. Notice that off-axis pixels in D and E will contribute to compressive-shear in some cases, and expansive-shear in other instances resulting in different amounts of motion blur. We have successfully used both approaches in different instances. The required computations are easily tractable with standard microprocessor based systems at several frames per second. More complex algorithms such as bundle adjustments are also possible, but may not be necessary for certain imaging platform altitudes.

4 Accommodating Dynamic Variations in Operational Camera Arrays Using Pose Information

The on-board global positioning system (GPS) sensor for estimating the aircraft position in world coordinates and on-board inertial measurement unit (IMU) sensors for measuring platform velocity, orientation and gravitational forces provide the necessary information to relate each computed WFOV image to be geo-registered in the context of a WGS84 world geodesic spatial coordinate system. The aircraft is in steady motion, and the camera array is under visual-servo control trying to maintain sight of a fixed patch on the ground. The servo-control system, also known as a gimbal-steering system, exhibits a finite delay. In addition, the GPS sensor can suffer spurious noise from time to time and the IMU has both drift error and shot noise associated with measurements. These dynamic variations and uncertainty in platform and camera-array orientation result in frame-to-frame jitter. Image jitter can be compensated locally (i.e., within a single-camera view) using image stabilization techniques often employed in other video applications [26]. Globally compensating for image jitter which occurs across the camera-array image planes requires the development of new techniques.

A more challenging variation is dynamical changes to the platform orbital path during different periods of time and lack of reliable camera-array position and orientation information. That is, environmental constraints during specific imaging-missions may necessitate non-circular (i.e., elliptical, zig-zag, criss-cross) flight-path trajectories of the aircraft and/or loss of camera-array navigation information. In such scenarios it may still be possible to analyze the acquired WAMI in a limited fashion depending on the actual platform trajectory and metadata reliability. In cases where navigation information is available and the flight path can be isomorphically remapped to a circular trajectory that also maintains temporal ordering of the views, then a new persistent WAMI sequence (nearly) equivalent to that from a circular orbit could be synthesized. In other situations where sufficiently accurate navigation information is not available, then temporal ordering may be relaxed and only the most consistent set of poses extracted. In such scenarios it is desirable to analyze each computed WFOV image and reorder the video sequence based on the relative pose associated with the observation of several structures of interest on the ground. Preliminary work in this direction has been investigated using experimental image sequences [16] that can be extended to very large format, wide-area WFOV imaging systems. Principal component analysis (PCA) is one potential tool for reordering a sequence of images by pose as described next.

Suppose each image $\mathbf{x}_i$ of size $M \times N$ of a randomly captured sequence of images is formed into an $MN \times 1$ vector. Suppose this is done for the entire set of images and the resulting vectors are made the columns of a matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_K]$ of random object poses. Then, the range space of the matrix $\mathbf{E}$ with columns composed of the L largest corresponding eigenvalues for some L is an L-dimensional subspace of R^{MN}. We will refer to this subspace as the *eigen-subspace*. The projection $\mathbf{g}(\mathbf{x}_i)$, of size $L \times 1$, of the image $\mathbf{x}_i$, onto the L-dimensional largest variance eigen-subspace is given by the expression,

$$\mathbf{g}(\mathbf{x}_i) = \mathbf{x}_i^T \mathbf{E}, \tag{12}$$

which is used as the associated feature vector for pose reordering.

The approach for reordering the images based on this feature is an iterative process. Let S_j and $\mathbb{S}_j$ be the set of unordered and ordered images at iteration j, respectively. To begin, S_0 is the entire set of unordered images and $\mathbb{S}_0$ is the empty set of ordered images. At iteration $j = 1$, a randomly chosen image is labeled $\mathbf{x}_1$ and moved from S_0 to $\mathbb{S}_0$, yielding S_1 and $\mathbb{S}_1$. For $j \geq 2$ an image $\mathbf{x}_j$ is moved from S_{j-1} to $\mathbb{S}_{j-1}$ such that,

$$\mathbf{x}_j = \underset{\mathbf{x} \in S_{j-1}}{\operatorname{argmin}} \big(\big\| \mathbf{g}(\mathbf{x}_{j-1}) - \mathbf{g}(\mathbf{x}) \big\| \big). \tag{13}$$

Thus, the ordering algorithm picks from the unordered set, the image closest to the last ordered image in the eigen-subspace. Once the images have been ordered using the minimum separation, a confidence measure is computed using local curvature along the trajectory (called the object manifold) of the ordered images in the eigen-

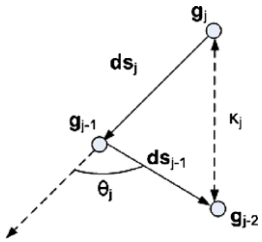

Fig. 11 Curvature explanation in a 3D eigen-subspace

subspace. Let $\mathbf{ds}_j$ be the vector, $\mathbf{ds}_j = \mathbf{g}(\mathbf{x}_j) - \mathbf{g}(\mathbf{x}_{j-1})$, then the cosine of the angle between the vectors is the correlation coefficient,

$$\cos(\theta_j) = \frac{\mathbf{ds}_j^T \mathbf{ds}_{j-1}}{\|\mathbf{ds}_j\| \|\mathbf{ds}_{j-1}\|}. \tag{14}$$

The magnitude of the amount of change between two adjacent difference vectors or three ordered image vectors, as illustrated in Fig. 11, can be computed as

$$\kappa_j = \|\mathbf{ds}_{j-1} - \mathbf{ds}_j\|, \tag{15}$$

which is a second order derivative approximation for the local manifold curvature. The confidence in ordering metric is given by

$$c_j = \exp\big(-\kappa_j\big(1 - \cos(\theta_j)\big)\big). \tag{16}$$

The confidence metric c_j attempts to use a combination of three local image projections to measure the alignment and the curvature. The alignment is equivalent to the congruence coefficient across three images and is equal to one when they are in a straight line. The curvature acts as a weight across the combination of the three images. A high confidence measure indicates the images are changing slowly and pose ordering is more accurate in this region, and a low measure of confidence means that the images are changing more erratically. A flow chart showing the PCA approach for pose ordering is shown in Fig. 12.

5 Summary and Conclusions

Wide-area persistent airborne video is an emerging very large format video with a specialized optical design for capturing large aperture images using an array of cameras. A well configured geometric arrangement is important to enable efficient image remapping, registration and mosaicking for producing accurate, extremely large 16K $\times$ 16K images at several frames per second sampling rate. The airborne camera array enables a denser sampling of the 4D light field in urban environments at higher spatial and temporal resolution than previously possible using other optical systems such as satellites or distributed single-camera airborne systems. Wide-area motion imagery, once it becomes more widely available, will facilitate the development of a new class of computational vision applications including dense 3D reconstruction of urban environments, continuous monitoring of large geographical

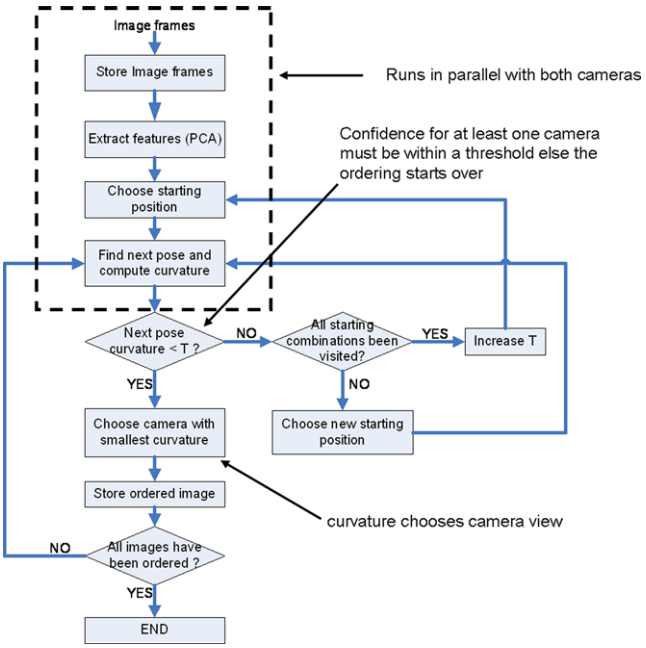

Fig. 12 Flow chart for multi-camera pose ordering algorithm

areas to analyze human activity and events, and surveys of large spatial regions to provide situation awareness for both civilian and defense needs. The optical geometry of an airborne camera array and a broadly applicable wide aperture virtual focal plane imaging model was developed to understand the image-sensor characteristics. Practical challenges in developing and using wide area imagery were described as a guide towards improving the utility of future systems. Some of the challenges in the exploitation of wide-area airborne video include the need for improved camera calibration, better estimation of platform dynamics, accurately modeling the spatio-temporal variability of the reflectance function across the camera array, and seamless image mosaicking. Given that the volume of data that can be captured by even modest wide area sensors is on the order of several terabytes per hour, or two orders of magnitude higher than standard definition video data rates, there is a pressing need for scalable on-board processing and tools to manipulate such large data sets for interactive visualization and analysis. A strategic research direction is multi-core vision algorithms for close-to-the-sensor processing to provide realtime geo-registration, compression, feature extraction, image matching, mosaicking, and object detection. Developing higher level algorithms for automatic 3D reconstruction, object tracking, and activity analysis offers additional research directions over the next decade. Exploring the parallelization of such algorithms across heterogenous computing systems will be critical to enable the timely use of wide-area large format video sensor data.

Acknowledgements The authors wish to thank Dr. Ross McNutt of PSS for providing the wide-area imagery used in this paper, Dr. Filiz Bunyak for various discussions and producing the figures related to spatio-temporal reflectance variations, and Joshua Fraser for creating the Maya-based rendering of the airborne imaging platform flight path and geometry. A new version of the Kolam software tool to support visualization of wide-area airborne video was developed by Joshua Fraser and Anoop Haridas and used for preparing the figures showing imagery in the paper. Matlab mex files to access PSS imagery especially for tracking was contributed by Ilker Ersoy as well as managing the collection of WAMI data sets. This research was partially supported by grants from the Leonard Wood Institute (LWI 181223) in cooperation with the U.S. Army Research Laboratory (ARL) under Cooperative Agreement Number W911NF-07-2-0062, and the U.S. Air Force Research Laboratory (AFRL) under agreements FA8750-09-2-0198, FA8750-10-1-0182. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of LWI, ARL, AFRL or the U.S. Government. This document has been cleared for public release under case number 88ABW-2010-2725. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

References

1. Adelson, E.H., Bergen, J.R.: The plenoptic function and elements of early vision. In: Landy, M., Movshon, J.A. (eds.) Computational Models of Visual Processing, pp. 3–20. MIT Press, Cambridge (1991)
2. Andrienko, G., Roberts, J.C., Weaver, C. (eds.): 5th Int. Conf. Coordinated & Multiple Views in Exploratory Visualization (2007)
3. Bunyak, F., Palaniappan, K., Nath, S.K., Seetharaman, G.: Fux tensor constrained geodesic active contours with sensor fusion for persistent object tracking. J. Multimed. **2**(4), 20–33 (2007)
4. Bunyak, F., Palaniappan, K., Nath, S.K., Seetharaman, G.: Geodesic active contour based fusion of visible and infrared video for persistent object tracking. In: 8th IEEE Workshop Applications of Computer Vision (WACV 2007), Austin, TX (2007)
5. Chou, E.C., Iyengar, S.S., Seetharaman, G., Holyer, J., Lybanon, M.: Velocity vectors for features of sequential oceanographic images. IEEE Trans. Geosci. Remote Sens. **36**(3), 985–998 (1998)
6. Cornbleet, S.: Geometrical optics reviewed: A new light on an old subject. Proc. IEEE **71**(4), 471–502 (1983)
7. Easson, G., DeLozier, S., Momm, H.G.: Estimating speed and direction of small dynamic targets through optical satellite imaging. Remote Sens. **2**, 1331–1347 (2010)
8. Ertl, T.: Guest editor's introduction: Special section on the IEEE symposium on visual analytics science and technology (vast). IEEE Trans. Vis. Comput. Graph. **16**(2), 177 (2010)
9. Fennell, M.T., Wishner, R.P.: Battlefield awareness via synergistic sar and mti exploitation. IEEE AES Syst. Mag. 39–45 (1998)
10. Hafiane, A., Palaniappan, K., Seetharaman, G.: UAV-video registration using block-based features. In: IEEE Int. Geoscience and Remote Sensing Symposium, vol. II, pp. 1104–1107 (2008)
11. Hafiane, A., Seetharaman, G., Palaniappan, K., Zavidovique, B.: Rotationally invariant hashing of median patterns for texture classification. In: Lecture Notes in Computer Science (ICIAR), vol. 5112, p. 619 (2008)
12. Hasler, A.F., Palaniappan, K., Manyin, M., Dodge, J.: A high performance interactive image spreadsheet (IISS). Comput. Phys. **8**(4), 325–342 (1994)
13. Hinz, S., Lenhart, D., Leitloff, J.: Detection and tracking of vehicles in low framerate aerial image sequences. In: Proc. Workshop on High-Resolution Earth Imaging for Geo-Spatial Information, Hannover, Germany (2007)

14. Kuthirummal, S., Nayar, S.K.: Multiview radial catadioptric imaging for scene capture. ACM Trans. Graph. (SIGGRAPH) **25**(3), 916–923 (2006)
15. Levoy, M.: Light fields and computational imaging. IEEE Comput. 46–55 (2006)
16. Massaro, J., Rao, R.M.: Ordering random object poses. In: IEEE Int. Conf. Acoustics, Speech and Signal Proc., pp. 1365–1368 (2009)
17. Nath, S.K., Palaniappan, K.: Adaptive robust structure tensors for orientation estimation and image segmentation. In: Lecture Notes in Computer Science (ISVC), vol. 3804, pp. 445–453 (2005)
18. Nayar, S.K.: Computational cameras: Redefining the image. IEEE Comput. Mag., Special Issue on Computational Photography, 30–38 (2006)
19. Nayar, S.K., Branzoi, V., Boult, T.E.: Programmable imaging: Towards a flexible camera. Int. J. Comput. Vis. **70**(1), 7–22 (2006)
20. Ni, K., Bresson, X., Chan, T., Esedoglu, S.: Local histogram based segmentation using the Wasserstein distance. Int. J. Comput. Vis. **84**, 97–111 (2009)
21. Palaniappan, K., Fraser, J.: Multiresolution tiling for interactive viewing of large datasets. In: Int. Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography and Hydrology, pp. 318–323. American Meteorological Society, Boston (2001)
22. Palaniappan, K., Uhlmann, J., Li, D.: Extensor based image interpolation. In: IEEE Int. Conf. Image Processing, vol. 2, pp. 945–948 (2003)
23. Palaniappan, K., Jiang, H.S., Baskin, T.I.: Non-rigid motion estimation using the robust tensor method. In: IEEE CVPR Workshop on Articulated and Nonrigid Motion, vol. 1, pp. 25–33, Washington DC, USA (2004)
24. Palaniappan, K., Ersoy, I., Nath, S.K.: Moving object segmentation using the flux tensor for biological video microscopy. In: Lecture Notes in Computer Science (PCM), vol. 4810, pp. 483–493 (2007)
25. Palaniappan, K., Bunyak, F., Kumar, P., Ersoy, I., Jaeger, S., Ganguli, K., Haridas, A., Fraser, J., Rao, R., Seetharaman, G.: Efficient feature extraction and likelihood fusion for vehicle tracking in low frame rate airborne video. In: 13th Int. Conf. Information Fusion (2010)
26. Ramachandran, M., Veeraraghavan, A., Chellappa, R.: Video stabilization and mosaicing. In: Bovik, A. (ed.) The Essential Guide to Video Processing, 2nd edn., pp. 109–138. Academic Press, Elsevier, New York (2008)
27. Regazzoni, C.S., Cavallaro, A., Porikli, F.: Video tracking in complex scenes for surveillance applications. EURASIP J. Image Video Process. (Special Issue), 1–2 (2008)
28. Rosenbaum, D., Kurz, F., Thomas, U., Suri, S., Reinartz, P.: Towards automatic near real-time traffic monitoring with an airborne wide angle camera system. Eur. Transp. Res. Rev. **1**, 11–21 (2009)
29. Sankaranarayanan, A.C., Veeraraghavan, A., Chellappa, R.: Object detection, tracking and recognition for multiple smart cameras. Proc. IEEE **96**(10), 1606–1624 (2008)
30. Seetharaman, G.: Three dimensional perception of image sequences. In: Young, T.Y. (ed.) Handbook of Computer Vision. Academic Press, San Diego (1994)
31. Seetharaman, G., Bao, H., Shivaram, G.: Calibration of camera parameters using vanishing points. J. Franklin Inst. **331**(5), 555–585 (1994)
32. Seetharaman, G., Gasperas, G., Palaniappan, K.: A piecewise affine model for image registration in 3-D motion analysis. In: IEEE Int. Conf. Image Processing, pp. 561–564 (2000)
33. Szeliski, R.: Image alignment and stitching. In: Paragios, N. (ed.) Handbook of Mathematical Models in Computer Vision, pp. 273–292. Springer, Berlin (2005)
34. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer, Berlin (2010)
35. Weng, J., Huang, T.S., Ahuja, N.: Motion and Structure from Image Sequences. Springer, Berlin (1991)
36. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. **38**(4), A13 (2006)

37. Yue, Z., Guarino, D., Chellappa, R.: Moving object verification in airborne video sequences. IEEE Trans. Circuits Syst. Video Technol. **19**(1), 77–89 (2009)
38. Zhou, L., Kambhamettu, C., Goldgof, D., Palaniappan, K., Hasler, A.F.: Tracking non-rigid motion and structure from 2D satellite cloud images without correspondences. IEEE Trans. Pattern Anal. Mach. Intell. **23**(11), 1330–1336 (2001)