2014

Automated Complexity-Sensitive Image Fusion

Brian Patrick Jackson
Wright State University

AUTOMATED COMPLEXITY-SENSITIVE IMAGE FUSION

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

By

BRIAN JACKSON
B.S.C.S., Wright State University, 2010

2014
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

December 3, 2014

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Brian Jackson ENTITLED Automated Complexity-Sensitive Image Fusion BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Arthur A. Goshtasby, Ph.D.
Dissertation Director,
Dir. CSE Ph.D. Program

Robert E. W. Fyffe, Ph.D.
Vice President for Research and
Dean of the Graduate School

Committee on
Final Examination

Arthur A. Goshtasby, Ph.D.

Jack Jean, Ph.D.

Thomas Wischgoll, Ph.D.

Lang Hong, Ph.D.

Vincent Schmidt, Ph.D.

ABSTRACT

Jackson, Brian. Ph.D., Department of Computer Science and Engineering, Wright State University, 2014. Automated Complexity-Sensitive Image Fusion.

To construct a complete representation of a scene with environmental obstacles such as fog, smoke, darkness, or textural homogeneity, multisensor video streams captured in diferent modalities are considered. A computational method for automatically fusing multimodal image streams into a highly informative and unified stream is proposed. The method consists of the following steps:

1. Image registration is performed to align video frames in the visible band over time, adapting to the nonplanarity of the scene by automatically subdividing the image domain into regions approximating planar patches
2. Wavelet coefficients are computed for each of the input frames in each modality
3. Corresponding regions and points are compared using spatial and temporal information across various scales
4. Decision rules based on the results of multimodal image analysis are used to combine the wavelet coefficients from different modalities
5. The combined wavelet coefficients are inverted to produce an output frame containing useful information gathered from the available modalities

Experiments show that the proposed system is capable of producing fused output containing the characteristics of color visible-spectrum imagery while adding information exclusive to infrared imagery, with attractive visual and informational properties.

Contents

Contents **iv**

List of Figures **v**

1 Introduction **1**

1.1 Motivation 1

1.2 Problem Statement 2

1.3 Registration 2

1.4 Fusion 5

1.5 Conventions 7

2 Literature Review **9**

2.1 Registration and Multi-View Methods 9

2.2 Statistical Image and Region Analysis 12

2.3 Wavelet Analysis and Fusion 21

3 Proposed Approach **25**

3.1 Data Collection 25

3.2 Scene Complexity-Adaptive Hierarchical Registration 27

3.3 Spatiotemporal Analysis and Decision Process 44

3.4 Local Wavelet-Domain Fusion 52

3.5 Quality Measurement 54

4 Results **61**

4.1 Registration 61

4.2 Analysis 66

4.3 Fusion 72

4.4 Quality Measurement 77

5 Significance and Contributions **83**

6 Summary and Conclusions **85**

6.1 Future Work 86

References **87**

List of Figures

3.1 The proposed system 25
3.2 Frames from the Calamityville datset 28
3.3 Frames from the PAMView dataset 29
3.4 The registration subsystem 30
3.5 Adaptive registration procedure 32
3.6 The Laplacian of Gaussian filter 33
3.7 Finding local extrema 35
3.8 Unsuccessful and successful transformations during RANSAC 38
3.9 Subdivision rules in the registration method 42
3.10 The analysis subsystem 44
3.11 Integral weighted motion 49
3.12 Three frame differencing for motion analysis 51
3.13 The fusion subsystem 52
3.14 Information shared by images b and f, but not a 60
4.1 Feature-rich and sparse areas in PAMView 62
4.2 Challenging regions for landmark correspondence 62
4.3 Bailout mechanism in the adaptive subdivision registration method 64
4.4 Results of the adaptive subdivision registration method 65
4.5 Interactive coregistration of EO and IR videos 67
4.6 Preprocessing in the IR band 69
4.7 Various intervals in IWM 70
4.8 Comparison of motion captured by IWM in two modalities 71
4.9 Selection of the variable δ for motion analysis 72
4.10 Comparison of motion captured by wavelet-domain three frame differencing 73
4.11 HSL decomposition of a color frame 75
4.12 Fusion results from the proposed method 76
4.13 Fusion results with additional distortion 77
4.14 Fusion results compared to other methods 78
4.15 Fusion results with artificial smoke effect 79
4.16 A closer view of fusion results across methods 80
4.17 $Q(I_0, I_1, F(I_0, I_1))$ 81

4.18 $Q(I_0, I_1, I_0)$ and $Q(I_0, I_1, I_1)$ compared to fusion output 81
4.19 $Q_A(I_0, I_1, F(I_0, I_1))$ 83

ACKNOWLEDGEMENTS

There are too many people to thank for their help in the production of this document.

I would like to thank the Air Force Research Laboratory and DAGSI for providing support and funding throughout the last four years. Particularly, I'd like to thank Dr. Vincent Schmidt, whose early, contagious excitement sparked my interest in image fusion, and Dr. John Camp, who not only set an example for those of us he left behind when he graduated, but also participated in the collection of the data featured in this work. I'd also like to thank Liz Pulley at Wright State, Bud McCormick at Calamityville, and Nicholas Gale and Evelyn Boettcher from Wright-Patt AFB, for making the video capture experiment at Calamityville possible, as well as the volunteers for their participation.

I would like to thank the members of my dissertation committee for participating in both the proposal and defense process, for their candid advice, and for their ongoing work with the students at Wright State. I'm very, very grateful to the Wittenberg University Math and Computer Science department, who counted me as one of their own even when I had so much left to learn. I'd also like to thank Philip Bohun, who was always there to hear my ideas before, during, and after their development.

I would especially like to thank Dr. Goshtasby, who has been the constant father-figure throughout my education. Dr. Goshtasby's mentorship and guidance has been an irreplaceable part of my career, and I owe him a tremendous debt.

Finally, I'd like to thank my wife, Emahlea Jackson, for being with me through thick and thin, through late nights writing, and early mornings working, and endless cups of coffee.

1 Introduction

1.1 Motivation

It is difficult for a single camera to capture a complete representation of a scene. Traditional electro-optical sensors cannot distinguish objects obscured by smoke or darkness. Infrared cameras provide thermal information about objects, but provide little information about the structure of background scenery at a uniform temperature. By constructing a system with multiple sensors, it is possible to overcome environmental obstacles such as fog, smoke, darkness, or homogeneity in one domain or another, enabling an observer to react to cues in any one of the video streams. The increasing availability of sensor technology to extend natural eyesight into other spectra has made multimodal camera systems attractive for a number of different applications. In recent years, firefighters have been increasingly equipped with thermal imaging cameras used to visualize areas of heat when entering burning buildings.Similarly, infrared and night vision technologies are often used to extend wilderness rescue [1], law enforcement [2], and wartime awareness of situations involving human presence.

Yet an abundance of information does not tend to simplify the task of an analyst. Though adding cameras to a system may mitigate the system's limitations, it also increases the complexity of the decision-making process. With a limited time to react to details displayed on screens, and limited attention to divide among them, the user's situational awareness depends greatly on how that information is presented. Image fusion, the process of combining information from multiple input images into an output emphasizing desirable properties, is a solution to the problem of growing redundancy and complexity in multimodal imagery. By giving a human operator imagery tailored to his or her task, useless information can be discarded, the operator's task can be simplified, and the most relevant details can be made plainly visible.

The benefits of simplification are especially evident in the operation of unmanned aerial vehicles (UAVs). Currently, multiple people may be required to operate one vehicle, and the analyst reviewing UAV footage is rarely in direct control of camera selection or flight path. By automating a portion of the selection process, the workload may be redistributed in favor of executive decision, saving manpower, decreasing the delay between observation and reaction, and reducing the amount of missed human action in the video streams.

1.2 Problem Statement

A camera rig mounted on an aerial platform captures frames in multiple modalities, including visible spectra and infrared. As the images are captured, both scene geometry and camera pose are subject to change: the former, from the motion of foreground objects, and the latter, from the motion of the platform as it shifts and rotates with respect to the scene. Additionally, the scene complexity is subject to change as the camera pans across the landscape. During some time intervals, the frames captured by the camera rig may contain occlusion, multiple depth planes, foreground objects, and other sources of nonrigid geometry, whereas frames captured in other intervals may contain very simple projective geometry with few foreground objects and no occlusion. Across modalities, information content may differ significantly as a result of lighting conditions, interfering media such as fog or smoke, misleading textures and contours introduced by camouflage, and noise inherent in the imaging process. Useful information from all input modalities should be aggregated, including background structures and moving objects. By applying an automated analysis and fusion technique, a unified view of the scene will be produced containing the useful information gathered across all available modalities.

1.3 Registration

Before automated analysis is applied, geometric differences between sequential frames of video data can be eliminated to separate temporal changes in the input data from spatial changes due to camera and target motion.

Registration is the process of computing spatial correspondence between two sets of data. Its applications are numerous; registration in two dimensions is used to align photographs in image mosaicking or video stabilization, and registration in three dimensions is a necessary step to fuse multimodal medical images [3].Through registration, multiple datasets obtained from the same scene can be reconciled, allowing analysis even when the datasets contain differences in camera position due to motion, differences in representation from multiple data collection or imaging techniques, or differences in content when the scene is imaged repeatedly over time. In general, a registration algorithm operates between a reference coordinate space S_R, which will remain unchanged, and a target coordinate space S_T, which will be transformed to correspond to S_R. Registration, then, seeks to solve for the transformation function T in the equation

$$\forall \vec{x_T} \in S_T : \vec{x_R} = T(\vec{x_T}) \tag{1}$$

where $\vec{x_T}$ and $\vec{x_R}$ represent corresponding points in the datasets. Since registration is primarily applied to images in this document, the notation $T(I)$ is used to denote an image produced by sampling images according to the transformation function, so that $T(I)(\vec{x}) = I(T(\vec{x}))$.

Often, correspondence in images is defined in terms of representation; for example, a location $\vec{x_R}$ in a reference image may contain the pixel representation of the northwest corner of a rooftop, and a location $\vec{x_T}$ is considered to correspond if a target image contains a representation of the northwest corner of the same rooftop, even if the perspective is different from the reference image. Correspondence is not defined thus exclusively – it is possible to conceive of determining the phase difference between two signals as a registration problem regardless of the semantic value placed on them [4]. Some applications of registration are also used to relate collected data to an idealized set of information, such as a medical atlas or a topographical map (or, the electronically-created equivalent: the DEM, or digital elevation model).

Registration methods are employed to mitigate differences in coordinate space between images exhibiting change in camera pose. In this application, correspondence must be established between pairs of images based upon the representation they contain of scene geometry, and the quality of the transformation function T obtained from a pair of frames in this way is assessed primarily on the function's alignment not of individual values, but of semantic content such as the textures, edges, and corners recorded in the data [3].If camera pose is changing over time, as opposed to simultaneous collection of imagery from multiple sensors, then this qualification overlooks the tendency in image sequences for changes taking place in the scene. Therefore, objects in the scene can be divided into two categories: foreground objects, which will change position and configuration in the scene over time, and background objects, which are stationary and act as reliable cues for the geometry of the scene.One key difference in registration methods is by what mechanism their procedures identify background points to reliably align the data, while ignoring or mitigating the potential errors caused by foreground objects.

An image registration technique can be based upon global information, such as the coefficients produced by applying a Fourier transform to the image, or localized information, such as the centers of shapes with homogeneous textures detected by Laplacian-of-Gaussian filtering [5][3].Depending on the application, either global methods or local methods may be preferable, and in some cases, a registation technique making use of both global and local informations – a hybrid registration technique – may outperform methods using only one or the other. Some global methods are well-suited to hardware implementation. This is especially true of methods employing global transformations like the fast Fourier transform, although recent hardware developments have also popularized optimized

versions of local methods employing SIFT, SURF, and FAST feature detection.Global methods tend to have a low dimensionality and few degrees of freedom, making them suited for simple scenes with high similarity. As a result of considering the entire image domain, many global methods are resistant to localized noise (for example, the presence of small foreground objects in motion). Local methods, by comparison, are capable of a higher number of degrees of freedom. While localized noise, occlusion, or interference may make certain local information unreliable, the use of local information allows for a much more flexible system, able to deal with non-planar geometry in a scene. The identification of points of interest called landmarks in both the reference and target frames lends itself not only to a class of methods to calculate correspondence, but also a number of measures for assessing the accuracy of a transformation function.

In practice, it is difficult to imagine a purely local or global approach to image registration. Hybrid registration techniques employing both global and local information can be very effective in certain applications. Transformations obtained by global registration can be refined using landmark-based methodologies. Whereas correspondence with no *a priori* information can be difficult to establish between two sets of landmarks with outliers, a global approximation of the transformation function facilitates the matching of landmarks.Global methods do not always preceed local; a global transformation function can be built by assembling in a piecewise or interpolated fashion between many local transformations obtained from landmarks.

Image registration applications vary greatly depending on the characteristics introduced by camera configuration, and specifically, the type of camera pose encountered throughout the dataset being registered. Approaches formulated for aerial datasets where the camera is pointed downward (near the nadir angle) will not be successful if applied to datasets obtained at the ground level at a horizontal angle. Zoom motion causing increasing scale of approaching objects is not present in nadir angle imagery while the camera is moving parallel to the ground; different assumptions must be made about the scene geometry and the camera's relationship with scene objects. In automated driver-assistance systems, increasingly being developed for high-end automobiles, one can reasonably assume that the lower portion of an image frame will contain a representation of the ground plane such that an object's lower boundary indicates its depth in the scene [6]. By contrast, an aerial scene implies that the majority of the scene is on the ground plane, and therefore that the background regions are well-registered with a single transformation function, sometimes as simple as a single projective transformation.

The difficulties associated with different types of imagery also change based upon the camera angle. Nadir-angle video, in particular, carries many difficulties: building surfaces occlude not only

the ground plane, but each other. In urban settings, the geometry can be very complicated in both the foreground and background; the elevation of the camera and the complexity of the topography in the scene makes it possible to observe a non-planar background, violating an otherwise helpful assumption that the world is flat. Horizon-angle video rarely sees enough terrain at one time to violate its assumptions concerning a flat, projective ground model. This is not to say *none* of the weaknesses present in nadir-angle video are present in other camera configurations. Regardless of camera pose, it can be difficult to relate significant contours or features in the image to the geometry of the scene. Regardless of camera pose, a registration algorithm operating on long sequences of images may encounter many types of surroundings in varying levels of complexity. Both local and global changes over time are a reality in nearly any camera configuration.

To mitigate the changes in image geometry introduced by both camera and scene motion throughout a video sequence, an adaptive subdivision technique for image registration is proposed. The technique applies an affine registration based upon the consensus of feature points in a downsampled image pair, then hypothesizes a similar transformation in quadrants of progressively higher-scale images, merging or splitting the registration procedure as necessary, then applying the resultant nonrigid transformation to the original frames. This will produce a global transformation function deformed by local non-rigidity.

1.4 Fusion

Image fusion is the process of combining information from multiple images, yielding a single image containing desirable aspects taken from the inputs. In contrast to registration, which obtains geometric relationships between images, fusion determines how images will be combined and represented. Unlike registration, image fusion tends to be anthropocentric: whereas the success of registration is quantifiable in terms of correspondence and accuracy, fusion's primary goal is to facilitate a viewer's understanding of the data. In general, the process of image fusion is determining the fusion function F for the ordered set of input images $\mathbf{I}$ in a common coordinate space S, producing an output image $O(\vec{x})$ according to the equation

$$\forall \vec{x} \in S : O(\vec{x}) = F(\mathbf{I}, \vec{x}) \tag{2}$$

Fusion, especially between registered images, can be applied to merge imagery having different focus levels or different exposures (a common method for simulating high dynamic range photography) [3]. Fusion is also used to bring information from different modalities together, and is particularly useful when different modalities provide complementary information about a scene. In medicine,

one image may represent the rigid structure of a human skeleton while another represents its soft tissues.In surveillance, one image may contain range data while another contains visible texture of objects in the scene. The method of analysis used to guide fusion must be selected based upon the task at hand. Local methods emphasizing a property of interest may be chosen for certain tasks, whereas global methods indiscriminate of an individual region's desirability may be more appropriate for other tasks.

Often, the fusion algorithm will vary depending on the aspects of the input images to be emphasized or incorporated into the output image O, and as a result, F will differ greatly from application to application. For instance, multi-exposure fusion may require an algorithm maximizing the information content of different regions among the input images, whereas multi-focus fusion depends not on maximizing contrast (which might emphasize blurred regions), but rather on sharp edges, measured by gradient-based statistics. Depending on the problem, fusion may appeal to information theory, intensity or gradient measures, clustering properties, or frequency coefficients. Computing image statistics across the regions of the image is a common task in fusion algorithms, and selecting statistics that emphasize the desired properties of interest is fundamental to the construction of a fusion algorithm.

With multimodal fusion, "completeness" can be thought of as the goal. This consists of combining the aspects of the input imagery that are hidden in some modalities, but revealed in others. Particularly in surveillance, the use of one electro-optical camera is subject to interference: naturally occuring media such as fog and rain can conceal a target of interest, and smoke or camouflage may be employed by a target. Important details in the scene are obscured easily, and night-time observation only complicates the task. The addition of infrared cameras to provide night vision is common; targets that would be otherwise difficult for an operator of such a system to identify can be much more easily detected in other spectra under certain conditions. A complete understanding of the scene, then, depends on information provided from multiple streams of imagery available to the system, and the result of fusion applied to the imagery should contain the necessary details from the input images to give the same understanding of the scene in a unified form.

To fuse multimodal videos containing objects of interest, a novel analysis and fusion method is proposed. To describe the relevant details of a scene, a decision map based upon the presence of moving foreground objects and the structure of background scenery is constructed. The response of spatiotemporal analysis methods to distinguish background from foreground produces values in the decision map corresponding to potential targets for tracking and observation, while including heterogenous background scenery in the map when foreground motion is not present. The images are

brought into a common representational format by applying the decision map to wavelet coefficients of the input images, resulting in a set of coefficients that can be inverted to produce an output image retaining the useful information from each of the input image streams.

1.5 Conventions

In this document, points are represented as homogeneous column-vectors. That is, the Cartesian coordinate pair (x, y) is represented as

$$\begin{bmatrix} x' \\ y' \\ w \end{bmatrix} \tag{3}$$

with w as a scale factor, so that $x' = wx$ and $y' = wy$. Where possible, vectors representing points will be assumed to be normalized, i.e. $w = 1$ to allow for equivalence with Cartesian intuition. Thus, for the sake of simplicity, functions traditionally accepting a two-dimensional vector can be written $f(\vec{x})$ or $f(x, y)$ with the assumption that the appropriate normalization or conversion is trivially applied. Notably, operations such as magnitude or length refer to the Euclidian length of the vector, and as such, are invariant to changes in w.

Images will be represented as a function of position (e.g. $I(\vec{x})$), with image streams also receiving an argument for time (e.g. $I(\vec{x}, t)$) with the bounds in the spatial and temporal domains left implicit, and the scales representing pixels and frames unless otherwise stated. When appropriate, the arguments to such an image function may be omitted – that is, the results of a pointwise function in the intensity domain $f(v)$ applied to each pixel in an image domain of an arbitrary frame from $I(\vec{x}, t)$ may be written $f(I)$ with

$$\forall t \in f(I), \forall \vec{x} \in f(I) : f(I)(\vec{x}, t) = f(I(\vec{x}, t)) \tag{4}$$

If S is a pixel domain, such as a window or image boundary, the image created by only considering points within S is depicted $I|s$, with $I_0, I_1|S$ denoting two images used within a given context are both limited to the area of consideration S represents.

Color data in an image is represented as a triplet, with $I(\vec{x}) = \{r, g, b\}$ for traditional RGB imagery. If a function $f(v)$ is defined for grayscale intensity values, unless specified otherwise, the function can be applied on a per-channel basis, with

$$f(\{r, g, b\}) = \{f(r), f(g), f(b)\} \tag{5}$$

thus allowing per-channel, per-pixel application of f to a color image I to be written $f(I)$, the same as grayscale images. Decompositions such as the FFT, and later, wavelet transforms should also be interpreted in this fashion, with coefficients replaced by tuples for color imagery and all basic operations on those tuples taking place independently per channel.

2 Literature Review

2.1 Registration and Multi-View Methods

The geometric analysis of aerial video is well-explored in the literature. Nevertheless, well-established registration algorithms do not satisfactorily apply to the problem at hand; the use case demands a procedure that can be applied online and to sufficient accuracy for valid statistical analysis of image regions. Synthesis of a new registration technique draws from the concepts of many simple techniques, yet must mitigate the error caused by a narrowly-constrained transformation function obtained by registration with few degrees of freedom.

One common approach to registration in aerial videos is the determination of disparity, a point-to-point transformation function not explicitly bound to a single global relationship, but rather determined locally in a piecewise fashion. The survey of literature suggests that a potentially effective avenue of online registration may be the combination of simple registration techniques applied in a local fashion and treated as cues to a non-rigid disparity function, stored as a map of bijections from reference points to target points.

2.1.1 Registration of Translated and Rotated Images Using Finite Fourier Transforms

DeCastro and Morandi, in a seminal 1987 paper [4], suggested the use of finite Fourier transform-based, frequency-domain registration. This registration, a rigid registration, produces three degrees of freedom in two dimensions: a translation in two directions, and a rotation within the image plane. The result is a registration function of the form

$$T(\vec{x}) = \begin{bmatrix} \cos\theta & -\sin\theta & \Delta_x \\ \sin\theta & \cos\theta & \Delta_y \\ 0 & 0 & 1 \end{bmatrix} \vec{x} \tag{6}$$

Registration takes place in two steps. The first step determines translation via the Fourier Shift Theorem, which relates a rotation in the spatial domain to a phase change in the frequency domain. For two images I_0 and I_1, the fast Fourier transform (FFT) produces $\mathcal{F}(I_0)$ and $\mathcal{F}(I_1)$, two-dimensional matrices containing Fourier coefficients. The cross-power spectrum is computed, multiplying one FFT result with the complex conjugate of the other result according to the equation

$$R = \frac{\mathcal{F}(I_0)\mathcal{F}(I_1)^*}{|\mathcal{F}(I_0)\mathcal{F}(I_1)^*|} \tag{7}$$

The inverse Fourier transform $\mathcal{F}^{-1}(R)$ produces an image containing an impulse at the coordinates (Δ_x, Δ_y) representing the optimal translation parameters between I_0 and I_1. If no rotation is present in the images, this method adequately determines T for the two images. By itself, the method for estimating translation in this way is called phase correlation.

The second step, employed when images also exhibit a rotational difference, estimates the rotation parameter via the Fourier Rotation Theorem. To produce a rotational parameter, R can be searched for an optimal θ by changing the cross power spectrum's normalization coefficient to a rotated version of the product to produce a unity pulse at (Δ_x, Δ_y). The method has also been extended by Reddy and Chatterji to find some scale differences between images [7].

2.1.2 A survey of hierarchical non-linear medical image registration

Lester and Arridge [8] detailed a number of course-to-fine registration approaches in medicine, grouping them broadly into methods with hierarchical data complexity (that is, the information at each step is increased from some basic level), warp complexity (describing transformation of increasing complexity), and model complexity (with matching methods increasing in sophistication). Many methods, including Gaussian pyramids, spline warps, elastic models, and scale space transformations, are reviewed briefly, with strengths and difficulties given for each. Some features to be emphasized include data reduction during the initial phases of registration, and some pitfalls, such as corruption of data by repeated resampling, are to be avoided.

2.1.3 Multimodal Stereo Image Registration for Pedestrian Detection

Krotosky and Trivedi [9] developed a method for registering color and thermal images with differences in scene depth in the image plane. Their method, employing a sliding-window approach, produces disparity between the views by maximizing mutual information using a voting scheme for each pixel location. Making the assumption that homography between the objects of interest in the images can be reduced to a simple displacement, a standard joint probability mass function and mutual information function is computed for each possible displacement for the windows containing objects of interest, generating a disparity matrix giving the optimal displacement for each pixel and a confidence value for that displacement.

The procedure forms the disparity voting matrix $D(u, v, d)$ (or, equivalently, $D(\vec{x}, d)$), where d is a displacement. As the matrix is formed, an entry at $D(\vec{x}, d)$ is incremented when it receives a vote. Then the optimal displacement is given by $D^*(\vec{x}) = \underset{d}{\operatorname{argmax}}\, D(\vec{x}, d)$, and the confidence of that displacement is $C^*(\vec{x}) = \max D(\vec{x}, d)$

It follows that the final registration function can be represented in terms of D^* as

$$T(\vec{x}) = \vec{x} + D^*(\vec{x}) \tag{8}$$

The authors remark that displacement maps generated this way can be used to refine a segmentation process, especially in use cases with stereo vision observing pedestrians. While the results are appealing, there is no ground-truth analysis, and the comparisons made by the authors are largely visual.

2.1.4 A multi-view approach to motion and stereo

Szeliski [10] prioritized accuracy and resolution in his 1999 work, detailing a procedure for computing depth and motion estimates in video streams.

The first approach he mentions – building a disparity space to relate voxels to surfaces – is reminscent of Krotosky and Trivoli [9], in that aggregation of disparity evidence takes place, upon which a map can be obtained. Szeliski writes seven years prior and suggests a more complicated three-dimensional representation.

The second approach involves decomposing the image into layers related to the objects so that each layer's pixels move according to a parametric transformation. Planar algorithms are suggested (although, like Dai [11], the EM algorithm would be involved in optimizing the number of layers).

To reconcile the weaknesses of the two approaches with the limitations the author has imposed, the author suggests computing a depth / motion map with each input image and establishing a compatibility constraint, determining occlusion relationships by computing the visibilities of pixels from frame to frame. The use cases cited for such a system include view interpolation (generation of images from an existing collection of images and depth maps), motion-compensated frame interpolation (prediction of future and past frames that can be used in compression or video processing), and construction of a segmentation-friendly representation of the data.

The stereo matching problem contains three subproblems: computation of matching costs, collection of local evidence, and determination of disparity values for each pixel. The paper details many approaches to each subproblem. The disparity problem in two dimensions, as in [9], is clarified here. Occlusion is listed as a problem in dealing with stereo maps, and addressed variously by the different cited authors.

Szeliski moves from outlining the problems in the domain to his specific solutions. First, keyframes, perhaps from characteristic views, are chosen for the computations of motion and depth

estimates. Next, the motion model between neighboring keyframes is computed as a function of constant flow with uniform velocity and rigid body motion. This allows both motion of the camera and observation of a moving rigid object. The formulation of the motion model is similar to the plane plus parallax representation, but can be applied without a dominant planar motion. The algorithm assumes global parameters such as feature point correspondence and tracking have already been computed so that the camera's egomotion can be obtained.

When analyzing pixel correspondence, matching uses a penalty function suggested in one of the author's previous papers, called a contaminated Gaussian distribution – a mixture of Gaussian and uniform distributions for which standard deviation of the inlier process and probability of outliers are both parameters that can be adjusted. The author notes that, while this distribution works as an estimate of the compatibility between neighboring pixels, analysis of residuals of computed probability distributions between neighboring pixels or disparities would be a better choice.

The author's cost function to optimize is composed of three terms: brightness compatibility, flow compatibility, and flow smoothness. The brightness compatibility measures the similarity of colors weighted by a visiblity factor related to occlusion, while also including adjustible parameters for global bias and gain. The flow compatibility measures consistency between neighboring keyframes by relating observed acceleration to expected variance of the motion model. The flow smoothness deals with discontinuity in intensity change in the neighboring frames.

Estimation of parameters is in two stages: the initialization phase, in which keyframes are considered independently, and the constraint phase, in which flow compatibility must be preserved, and visibility factors are computed. These two phases, including matching, can take place in a multiresolution pyramid. Two approaches for applying hierarchical reasoning to the problem are given: the correlation-style search and Lucas-Kanade gradient descent. In either case, hypotheses are refined and updated in an iterative fashion for each keyframe and on each level of scale, course to fine.

2.2 Statistical Image and Region Analysis

The goal of analysis in this project is the guidance of a decision process: the quantification of "relevance," "usefulness," or "importance" is paramount to the success of the system's fusion routines. To that end, segmentation of objects and modeling of image layers are convenient paradigms for determining the relative value of input data across the various regions of the image domain.

Whether the task is object avoidance (in the case of vehicle-mounted pedestrian detection) or

object tracking, many features are shared across image and region analysis techniques. First, the classification of pixels or components is a necessary step in reducing search space. Second, the use of rapid, online methods is crucial to allow reaction to take place in real time and with minimal latency. Third, the use of simple algorithms can be justified with statistical theory. Finally, the image is frequently decomposed into components that can be passed on to a separate subsystem.

By examining the common elements and benefits of many recent and long-standing analysis techniques, a highly appropriate mechanism for guiding the decision-making process can be developed and integrated into the proposed system.

2.2.1 Adaptive background mixture models for real-time tracking

Stauffer and Grimson [12] developed a highly useful technique for classifying background and foreground pixels based upon the modeling pixel locations over time as a mixture of Gaussian random processes. Their method addresses cumulative errors and changes in the scene over time in an adaptive fashion, yet tolerates periodic changes in the background that simpler methods such as mean or mode background subtraction would not allow.

A pixel's color over time is represented as a "pixel process," a random process modeling each static color observed in a noisy environment as a single Gaussian. For any such pixel process, the history at a certain instant can be used to approximate a probability density function, from which the probability of observing the current pixel color can be derived. From this information, k Gaussian distributions are constructed for each pixel at each instant, so that observations for each pixel can be compared to the previously constructed distributions and matching it to the process that best predicts the value (or, if none of the previous processes are likely, replacing the least observed with a new process). Making the assumption that "background" processes for any given pixel are represented by the Gaussians with a high amount of supporting evidence from previous observations as well as a small standard deviation over time, the Gaussian processes can be classified into background and foreground processes. Hence, the observations matched to background or foreground processes are classified as background or foreground at a moment in time. A basic outline of the algorithm is given below:

For each frame $I(\vec{x}, t)$ to be processed, a background model is updated

$$M(\vec{x}, t) = \left\{\mu_i, \sigma_i^2, w_i | 0 \leq i < k\right\} \tag{9}$$

The model consists of k Gaussian distributions for each pixel, represented by a mean μ, variance

σ^2, and weight w. These distributions represent average colors of either foreground or background objects observed at $\vec{x}$ over some period of time, where distributions with high w represent those more frequently observed.

Each observed intensity is matched to one of the k previous distributions by finding the highest-weight distribution with a euclidean-norm distance of less than $2.5\sigma_i$. If none of the distributions match, the new value replaces the lowest-weight distribution, assuming that the new σ_i^2 is large, and the new w_i is small.

Assuming the matched distribution at $(\vec{x}, t)$ has parameters μ_j, σ_j^2 , and w_j, the model is updated with

$$\begin{aligned} \mu_j(\vec{x}, t) &= (1-\rho)\mu(\vec{x}, t-1) + \rho I(\vec{x}, t) \\ \sigma_j^2(\vec{x}, t) &= (1-\rho)\sigma_{t-1}^2 + \rho \|I(\vec{x}, t) - \mu_j(\vec{x}, t)\|^2 \\ w_j(\vec{x}, t) &= (1-\alpha)w_j(\vec{x}, t-1) + \alpha \end{aligned} \tag{10}$$

where α is a learning parameter controlling how quickly unobserved distributions decay over time. The parameter ρ is derived from α, with $\rho = \alpha\eta(I(\vec{x}, t)|\mu_i, \sigma_i)$.

The weights for all unmatching distributions are also updated with

$$w_j(\vec{x}, t) = (1-\alpha)w_j(\vec{x}, t-1) \tag{11}$$

before the weights for all distributions are normalized.

It remains, then, to determine which of the distributions represent the colorspace of background objects and which can be classified as foreground. At each $(\vec{x}, t)$, the distributions are sorted by w_i/σ_i, recognizing that backgrounds will be more static and more frequently observed at a given location than foreground. Then the strongest B distributions are classified as background, where

$$B(\vec{x}, t) = \underset{b}{\operatorname{argmin}} \left(\sum_{i=0}^{b-1} w_i > T \right) \tag{12}$$

and T is the expected ratio of background observations to frames.

While the segmentation ability of the model can be increased using connected components analysis and a working knowledge of the scene's characteristics, the system requires initialization time to correctly construct the Gaussian processes over the image domain, and the memory use of the system will be directly related to the parameter k, determining how many processes can be retained for each pixel at a time (with k usually determined experimentally, although [13] produced an automated method with comparable results). Obtaining optimal parameters related to the expected ratio of

background to foreground pixels in the scene or the amount of adaptive delay before weakly-observed processes are forgotten is also necessary.

2.2.2 Improved adaptive Gaussian mixture model for background subtraction

The major novelty introduced by Zivkovic [13] in this paper is a subsystem for selecting the number of background modes for a GMM. The novel portion of the work, namely, the number of modes for the background at a point in the image domain, uses a Lagrange multiplier to formulate a maximum likelihood estimate for the ownership values attached to the modes of the Gaussian mixture. This is a reformulation of the original GMM basis. The maximum a priori solution for the recurrence relationship is also reformulated. The contribution of this author is given in one equation by adding a single negative constant to the recurrence relationship for the ownership values, then renormalizing the weights so that their sum is fixed at unity.

This very minor adjustment to the GMM approach has a remarkably small effect on its accuracy, a fact to which the ROC curve and the author's own analysis testify. While the reformulation of the GMM mathematics is perhaps useful in understanding the basis for the ownership functions, the extension the author has offered to the GMM approach is negligible in importance and impact.

2.2.3 Background modeling and subtraction of dynamic scenes

Monnet et al. [14] take a statistical approach to dynamic background modeling. Recognizing the importance of Stauffer and Grimson's work with Gaussian Mixture Models, the authors criticize the performance of the GMM in dynamic, non-stationary scenes. Using ocean waves, waving trees, rain, and moving clouds as examples of videos with a spatiotemporal pattern of change, the authors offer a predictive model based on subspace signal analysis of the video as a time series. Two mechanisms are used: an incremental update technique and a method for replacing the models using an observation map.

The prediction model is autoregressive, similar to the GMM. The prediction mechanisms should be constrained to use the k latest images in the series. The process uses a spatial filter option to reduce the search space, yielding a set of features from pixels in the scene. Particularly, the authors use linear filters, but suggest that other filters, including wavelets, would be feasible. From the state space, PCA will be used to generate basis vectors. Over time, the predictive model based upon these basis vectors will be updated, and the basis vectors will be adjusted as new frames are processed.

The initial model is constructed from the last m frames by singular value decomposition. Afterwards, the model is revised using "exponential forgetting," which resembles the decay function of the

GMM. The basis vectors can be updated without a full recomputation of the principle components; the authors offer a method called Incremental PCA for such a situation. This method will depend on the amnesic mean of observations over time – a measure analogous to the traditional mean, but in which the weight of older observations is decreased with exponential decay. Then the basis vectors can be pulled towards the amnesic mean in a sequential fashion from first basis vector to last, so that the residues of projecting the mean along each basis vector are passed on to the next vector in the series.

The predictive model, then, can be constructed by solving a set of linear equations to yield their optimal parameters: the k-th order autoregressive model is overdefined by the states generated by the above process, so normal equations are used to solve the system, optimizing in a least squares fashion the error between the predictive model and the difference in observations between the current state and the previous state.

Using the described model, objects are poorly predicted, while backgrounds are well-predicted. The authors pivot from the study of background modeling only to a consideration of the types of change detectible in the signal: there is "structural" change, where pixel intensities will change in a region, and "motion characteristics" change, where the change in the temporal domain is unusual.

An error measure computed by estimating the Mahalanobis distance between the prediction and the observation at any time is offered as a detection measure, approximating the distance from the Gaussian based upon the PCA results to the observation. This distance can be used to measure the change in scene structure due to appearing or disappearing objects or change in color, i.e. structural changes. The side-effect of the PCA used by the authors is that the technique effectively considers relationships between pixels rather than their individual properties as in the GMM.

The change in motion characteristics is represented by the square of the L_2 norm of the difference between the predicted state and the observed state. The authors describe this distance as the measure of how information appears in a different temporal order than the background.

The computational complexity of finding the basis vectors across an entire scene is quite high (roughly cubic), and so the authors break the image down into independent blocks, for which the SVD determines both number of components and number of past images to consider. Even with this level of optimization, the algorithm still runs at 5 fps on their test machine in a very low-resolution video. Extrapolating from their result, performance is expected to be a problem even on modern hardware when applied to the resolutions of data available to us and coupled to the other subsystems involved in our project.

The performance of the algorithm is given as an ongoing work in the paper. It is possible that

the approaches conjectured by the authors (that is, nonlinear operators, more elaborate prediction, and neighborhood interdependency analysis) would improve the accuracy, but it is unlikely that higher complexity will address the cost of computing such a measure in real time.

2.2.4 Statistical background subtraction for a mobile observer

Hayman and Eklundh [15] describe a more general use of background modeling than what was developed by Stauffer and Grimson. A moving camera with pan, tilt, and shift is described, using a hierarchy of algorithms approach to build a background model while the system is online. Noticeably absent is a good method for addressing motion of the camera; only pan and tilt are directly addressed by the authors' solution.

The performance metrics chosen by the authors include false alarm rate and misdetection rate. Additional requirements are imposed: the algorithm must be statistically sound and capable of automatically adapting to noise levels instead of requiring a manual threshold setting.

Stauffer and Grimson's GMM technique is adopted and refined. The authors move from restating the GMM assumptions and presenting the approach to an extension of the method for use with an "active head." This merely adds registration to the GMM approach (by using a grid larger than the active screen area) and attempts to mitigate problems the registration adds. Problems in GMM plus registration identified by the authors include registration inaccuracy (resulting in pixel process values being sent to the wrong mixture of Gaussians), sub-pixel inaccuracy, and motion blur, all of which are acutely problematic in areas of high texture.

The addition of a noise term and the use of convolution with a filter kernel remove some of the problem of mixed pixels. The system deals with the false alarm rate, but causes the misdetection rate to increase.

Covered background detection for quick initialization is a second problem addressed by the system. The authors' solution uses an alternate weight function for early frames, but uncovers a greater problem, which the authors do not address: rotation causes sub-pixel inaccuracy that alter the variance of the GMMs, making the suggested system unsuitable for a camera with rotation.

2.2.5 Wallflower: principles and practice of background maintenance

In 1999, Toyama et al. [16] presented a list of problems encountered by background maintenance systems and develop a system aimed at addressing the problems. By comparing their system to eight other background subtraction algorithms, the authors derived important principles in designing background subtraction systems.

The authors list ten common problems in background subtraction, only seven of which are addressed in their comparison:

- Background objects may be moved, and after some period of time has passed, these objects should be reabsorbed into the background model
- Gradual illumination changes take place in the scene
- Sudden, global illumination changes take place in the scene
- Backgrounds may exhibit periodic or noisy fluctuations
- Foreground objects may be absorbed by the background model
- There may be an insufficient number of frames to train the background model
- Moving objects may be homogeneous in texture, color, or intensity, so that only their borders differ from frame to frame
- Foreground objects may become motionless for a long period of time
- Background objects may become foreground objects
- Shadows may be cast by a foreground object

The approach suggested by the authors is called Wallflower. Wallflower models the background on three levels of abstraction: pixel-level (similar to IWM or GMM), region-level (similar to connected-components analysis), and frame-level.

On the pixel level, Weiner filters are used to determine if a pixel is predicted by observed values at the same location an arbitrary number of frames (the authors use 50) in history. The system also maintains a history of predicted values - essentially, a smoothed version of the Weiner filter intended to eliminate corruption of history values by foreground motion. The system adapts during each new frame, and is kept if the prediction error is below an arbitrary threshold (the authors use 10%).

On the region level, the intersection of foreground-classified pixels over three frames and one foreground object is used to seed a region-growing algorithm based upon the histogram of values in connected foreground components.

On the frame level, multiple background models are used, and the model is chosen for which the smallest number of foreground pixels is determined at a given time. This trumps the region-level algorithm, which cannot accurately grow foreground regions when a large global intensity change

occurs. It is possible that this algorithm would be useful in switching between multiple stationary camera views, although this is not an application mentioned by the authors.

The methods compared to Wallflower are simple background subtraction from adjacent frames, mean and threshold, mean and covariance, Gaussian mixture model, normalized block correlation, temporal derivatives, Bayesian decision, Eigenbackgrounds, and linear prediction models. When applicable and necessary, the authors generalized gray-level methods to RGB using an L_1 norm and three applications of the algorithms in question, with Javed et al. developing a superior color generalization in [17]. Removal of small four-connected islands of eight pixels or fewer was also enforced across the board. A helpful numerical table of comparison results is given.

The principles given by the authors as a result of their evaluation of Wallflower are of particular interest:

- Semantic differentiation of foreground objects ought not be tied tightly to background maintenance (although higher-level analysis and object recognition results may be passed downwards into the background model)
- Finding foreground objects is a separate problem from determining if they are objects of interest
- Pixel-level "stationarity" must be well-defined, so that pixels satisfying such a measure can be definitively assigned a background class
- The background model must adapt to both sudden and gradual changes in the background
- Changes in the background model occur on different spatial scales (i.e. pixel scale *and* frame scale)

The authors admonish engineers to consider realistic and pragmatic goals such as the above. Tellingly, the comparison to certain already-established background modeling algorithms was favorable overall, but small changes in one or two areas put some of the techniques very close to Wallflower's performance; for example, the Gaussian mixture model is nearly equal to Wallflower in error rate except when sudden global illumination changes due to light switches occur, and the Eigenbackground method gives nearly equal performance except in cases where training is insufficient.

2.2.6 Layered Representation for Pedestrian Detection and Tracking in Infrared Imagery

The layered representation used by Dai et al. [11] (and later, in [18]) is of particular interest, bridging the gap between segmentation and decision maps used in fusion. In this representation, an image is modeled as the sum of three terms according to the equation $I_i = (1 - M_i)BG_i + M_iFG_i + W_i$, where an image's contents can be defined as a sum of the background (BG_i), the foreground (FG_i), and sensor noise (W_i), with a mask layer (M_i) carrying the critical information about objects of interest.

Dai et al. use an Expectation Maximization (EM) method to model the background of each frame, with a classification method to limit their segmentation to pedestrians, using Principle Components Analysis (PCA) on shape cues such as compactness and leanness to distinguish objects of interest. While an accurate segmentation is determined from this process, videos must be processed in a non-serial fashion to correctly analyze motion, and therefore, the algorithm is unsuited to real-time streaming imagery.

2.2.7 A Shape-Independent-Method for Pedestrian Detection with Far-Infrared-Images

Fang et al. [19] produced a pedestrian detection system using infrared cameras and a segmentation-classification process. Assuming that targets of interest exist on a horizontal plane, the segmentation method that is used separates the columns of the imagery into 1-D horizontal histogram-like measures. In each column of an image, the number of pixels above a threshold are counted. The number is stored in an array representing the "bright-pixel-vertical-projection curve." Such a curve has "bumps" separated by zero-level flat regions. The bumps represent potential areas of interest, and so the horizontal domain is split into interesting strips separated by irrelevant areas. Having split the image domain thus, a vertical segmentation is applied to produce rectangular ROIs for classification.

The two vertical segmentation methods proposed have different strengths. One is a brightness-based method helpful in sparse IR images (especially during winter or in suburban, uncluttered areas). The other is a body-ratio method. The brightness method finds, in any given stripe, the highest and lowest bright pixel. These denote the vertical bounds of the object. The bodyline method finds the locations in any given row of the most rapid intensity change: the left boundary being low-to-high and the right being high-to-low. The generated row lengths are fed to a histogram-based classification method, responsible for pedestrian detection.

Two pedestrian detection performance measures are proposed by the authors - they measure the

ability of the segmentation subsystem to select the entire pedestrian area (segmentation side accuracy) and the ability of the subsystem to select tight bounds on the pedestrian area (segmentation side efficiency). The methods proposed by Fang et al. for shrinking the search space for template matching are not applicable to far-infrared pedestrians - the symmetry property is not necessarily applicable to all poses, and the large number of poses in our project increase the complexity considerably.

2.3 Wavelet Analysis and Fusion

Wavelet analysis is a well-established method for determining properties of a signal originally applied in quantum mechanics and optics. Grossmann and Morlet [20] established that a family of wavelets, obtained by shifting and dilating square-integrable functions called "analyzing wavelets" or "mother wavelets," could be used to decompose another square-integrable function, forming an orthogonal basis in such a way as to make perfect reconstruction of the original signal possible given the decomposed (and irreducible) form.

A function $f(x)$ is considered square-integrable if it satisfies

$$\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty \tag{13}$$

The space of such functions, $L^2(R)$, contains all finite functions residing within a finite domain, as well as functions like Gaussians. A wavelet drawn from this space is a function that "wobbles" from the x-axis. Specifically, a wavelet function $\psi(x)$ satisfies

$$\int_{-\infty}^{\infty} \psi(x) dx = 0 \tag{14}$$

with mother wavelets usually satisfying the additional constraint

$$\int_{-\infty}^{\infty} |\psi(x)| dx = \int_{-\infty}^{\infty} |\psi(x)|^2 dx = 1 \tag{15}$$

Convolution of a wavelet with a signal in $L^2(R)$ produces high values when the wavelet resonates with the signal. Wavelet analysis, then, is concerned with the resonation produced by convolution with the wavelet at different scales (corresponding to different resolutions) and in different positions, from which information can be extracted about the signal at different frequencies.

2.3.1 A Theory for Multiresolution Signal Decomposition: The Wavelet Representation

Mallat [21] proposed the decomposition of images in signal processing using wavelets in a procedure that has come to be called the discrete wavelet transformation (DWT).

As a mother wavelet, the Haar wavelet is chosen, defined as

$$\psi(x) = \begin{cases} 1 & 0 \leq x < \frac{1}{2} \\ -1 & \frac{1}{2} \leq x < 1 \\ 0 & otherwise \end{cases} \tag{16}$$

The Haar wavelet, while discontinuous, forms an orthonormal basis when scaled by powers of two and applied, in the case of images, first to rows of image intensities and then to the columns of the result. A very common optimization of this procedure simply subtracts neighboring pixels and scales the image down by a factor of two (referred to by many of the following authors as the "decimation step"). Each application of the procedure produces four sets of coefficients in the image space: a set corresponding to horizontal edge components, a set corresponding to vertical edge components, a set corresponding to diagonal corners, and a downsampled version of the original image.

Regardless of the number of levels of decomposition produced by repeatedly applying the DWT to an image as the sampling rate decreases, the total number of coefficients is constant. The orthogonality of the wavelet series produces a complete decomposition without redundancy, often seen as a benefit in applications such as image coding and compression. The computational complexity of the DWT, in $\Theta(n \log n)$, is also attractive when dealing with large data or time-sensitive contexts. Nevertheless, the efficiency of the DWT comes with drawbacks: the coefficients produced by transforming two images with a small translation or scale are very different, hampering comparison between images separated by time or space. The directional components are also strongly aligned to the axes of the image frame, which makes transformation sensitive to small changes in rotation. These shortcomings continue to be addressed by specialized variations of the DWT, trading redundancy and increased computational requirements for greater flexibility.

2.3.2 A real-time algorithm for signal analysis with the help of the wavelet transform

Holschneider et al. [22] suggested the use of a similar wavelet approach, omitting the decimation step and producing an over-complete, stationary decomposition of the analyzed signal.

The stationary wavelet transform (SWT) or shift-invariant wavelet transform (SIWT) was orig-

inally given by Nason and Silvermann [23] as an undecimated form of the DWT. The algorithm à trous (French for "algorithm of holes") given by Holschneider et al. computes SWT coefficients in an image for increasing scales of the wavelet (or, equivalently, lower resolution of the analyzed signal) by constructing a 2D filter, and inserting zeroes into the kernel as the decomposition proceeds. By using a kernel with the same number of non-zero values spread out over the low-pass filtered image without downsampling the image, the transformation produces an image the same size as the original for each level of the transformation. Whereas the DWT applied to a n pixel image obtains n coefficients after k levels of decomposition, the SWT produces kn coefficients. Predictably, the increase in coefficients requires an increase in computation time required, although both the DWT and SWT are sufficiently efficient that real-time use is practical.

2.3.3 Motion Estimation Using a Complex-Valued Wavelet Transform

Magarey and Kingsbury [24] extended the DWT to the complex domain, formulating the complex discrete wavelet transform (CDWT). The use of a complex four-tap filter allows for simple integer-valued Gabor filter-like operations to replace the Haar wavelet basis of the simpler DWT. Instead of 0°, 45°, and 90° orientations, the wavelets are oriented at ±15°, ±45°, and ±75°, yielding six coefficient subimages instead of three for each level of decomposition and rejecting negative frequencies in each.

The attempt at formulating a complex, stationary extension of the DWT method was soon improved upon by Kingsbury. Unlike the DWT, the CDWT cannot obtain perfect reconstruction (due to the reconstruction block's inability to produce a flat overall frequency response, recognized by Kingsbury in [25]).

2.3.4 The Dual-Tree Complex Wavelet Transform: A New Technique for Shift Invarance and Directional Filters

Kingsbury's second 1998 treatment of the wavelet transform [25] addressed the shortcomings of the DWT and CDWT by introducing a parallel fully-decimated wavelet form, the dual-tree complex wavelet transform (DT-CWT), in which two sets of coefficients are produced by filters offset by half a sample. That is, one tree is formed from the odd components of the image, and the other is formed from the even. Kingsbury showed, by use of more sophisticated filters, that the Gabor-like filters similar to those originally used in the CDWT could be applied without the loss of perfect reconstruction.

Despite its tree structure, the DT-CWT is nearly shift invariant, similar to the SWT, demonstrated by observing the change in the energy of the different wavelet coefficients at each level of the transform as the signal being processed is gradually shifted. Lewis et al. [26] showed the feasibility of a DT-CWT segmentation algorithm, although coming short of a fully online system.

2.3.5 Shift Invariance in the Discrete Wavelet Transform

Bradley [27] generalized the space of algorithms between the DWT, which is sparse and shift variant, and the SWT, which is redundant and shift invariant. His work highlights the implicit tradeoffs between efficacy and efficiency of wavelet methods: the continuous wavelet transform (CWT) and the SWT are much more expensive to compute than the DWT or similar methods, but inverting the transformation to reconstruct the data is not as necessary in analysis.

To give a greater degree of flexibility, Bradley proposed a hybrid approach called the over complete discrete wavelet transform (OCDWT), in which certain high frequency levels of decomposition are critically subsampled, as in the DWT, while lower frequency levels are fully sampled, as in the SWT. Such an approach to wavelet transforms treats the DWT and the SWT as special cases of the OCDWT for which all levels are subsampled (producing the DWT) or which no levels are subsampled (producing the SWT). Additionally, Bradley asserts an OCDWT with one or two levels of decimation and critical subsampling followed by the SWT may still be sufficiently shift invariant to recover edges after shifting by a few pixels.

3 Proposed Approach

The proposed system (shown in Figure 3.1) consists of three software modules with their associated validation methods. First, registration will be applied, allowing analysis to take place in a consistent reference frame. Second, analysis will take place in spatial and temporal domains, producing a decision function and comparison measures for every pixel of each image for a given frame. Third, guided by the decision function, a local wavelet transform-based fusion method will be applied to selectively merge the images, giving preference to high-value regions from the input imagery and generating the output of the system. A quality measure will be employed to characterize the system's output and compare it to output of other fusion methods.

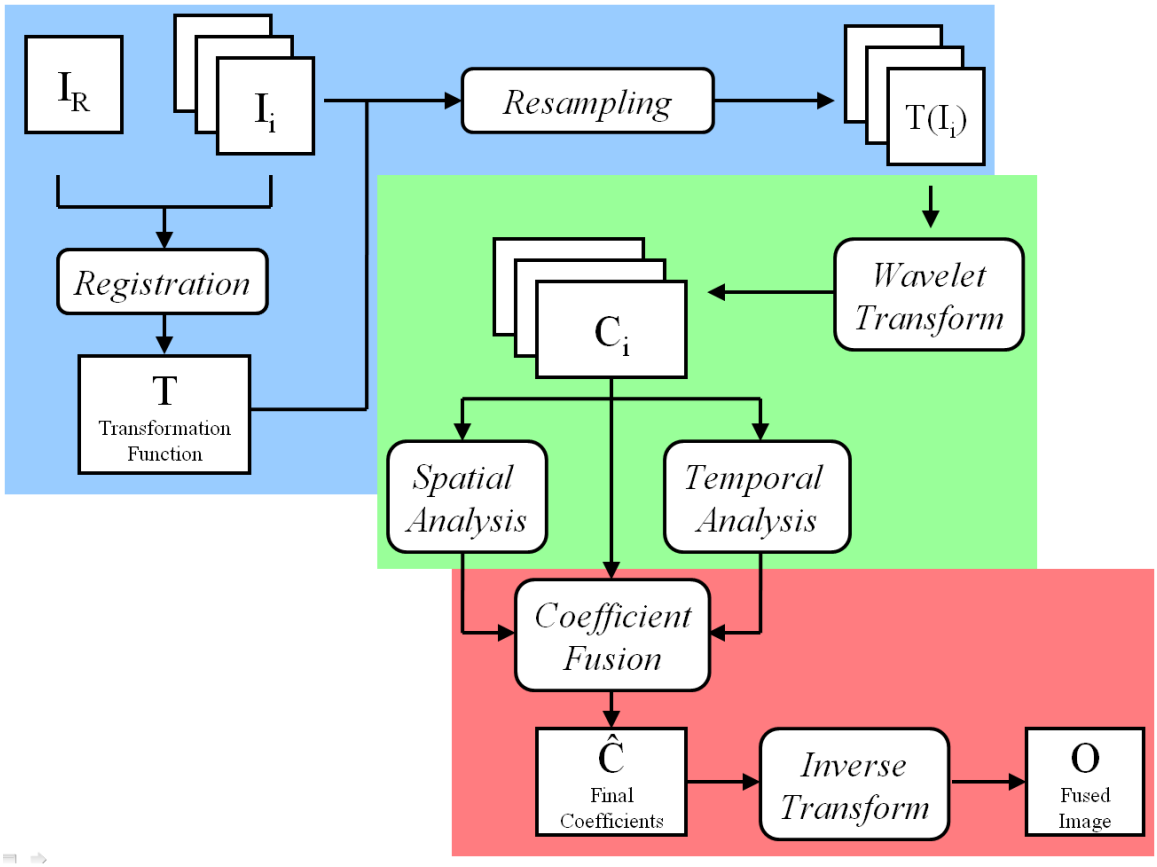


Figure 3.1: The proposed system

3.1 Data Collection

For each module in the system, there are requirements and constraints placed upon the videos to be used as input for testing purposes. For registration, the camera pose must be sufficiently close to the

nadir angle to avoid the degeneracies a horizontal orientation would cause; modeling the scene as globally planar causes regions near the horizon to be subdivided to a much greater degree than the rest of the image. For analysis, data must be synchronized and co-registered, so that the semantic details of objects in the various input modalities are identical. For fusion, the data must have objects and regions of interest, subject to change over time.

A new collection of multimodal imagery was proposed to will satisfy the following conditions:

- Each video will consist of a sufficient number of frames for analysis. This will be defined as no fewer than 500 frames, to allow for an initialization phase and an analysis phase to both take place.
- Each video will be sampled at a sufficiently high rate. This will be defined as no less than 4Hz, preferring videos closer to 24Hz or 60Hz. As the intended use case for the system is real-time analysis of streaming video, the sampling rate must allow for an observer to track objects in motion.
- Each video will consist of one or more infrared stream and one or more electro-optical stream. The negotiation of very similar spectra is not as interesting or helpful as choice between dissimilar imagery.
- The videos must contain some camera motion. Camera jitter introducing a random, but small amount of motion is assumed, but motion representing a camera path is desirable.

A series of data collections was carried out at the Calamityville facility in Fairborn, Ohio. The facility was well-suited to producing videos appropriate for demonstrating the proposed techniques. After a tour of the facility, it was clear that the outdoor training scenarios already constructed for disaster preparedness and search-and-rescue could be easily repurposed for the creation of relevant video containing human activity in a realistic environment. Using a vantage point over 100' in the air, a series of videos were recorded at a downward angle of approximately 45° representing a variety of scenarios. Each scenario was captured by three cameras: two Basler Ace acA2000-50gc cameras at 1920×1080 resolution and a FLIR T450sc camera at 320×240 resolution. The left Basler camera was configured to capture near-infrared imagery through a zoom lens, but the videos produced were considerably problematic for fusion: their color was inconsistent, and the regions of interest were too small to contribute meaningfully to the information of the scene. The right Basler camera, with a narrow focus (24mm, 24° field of vision), captured high quality electro-optical (EO) video, and the FLIR camera captured long wave infrared (LWIR) in the 7.5–13μm range. All videos were captured

at 30Hz, reducing the difficulty of synchronization, but the FLIR camera was controlled by hand, introducing temporal misalignment to the video, as well as minor jitter at the beginning and end of the videos.

The scenarios recorded made use of a combination of environmental and behavioral modifiers. Smoke machines were available to simulate occluding media, and several videos were captured with and without smoke. Additionally, rotational camera motion could be introduced, producing videos with stationary imagery, deliberate and gradual motion, and jitter. Finally, the subjects in the videos were instructed to carry out a variety of different actions: they were to be absent in some videos, stationary in one set, walking and performing non-threatening actions in another, and actively running to conceal themselves in others. The array of 14 scenarios was designed to be interesting to this project, while also allowing for future use in fusion, registration, and human detection studies. Frames from the various scenarios are shown in Figure 3.2

A second dataset was also used extensively during testing of the registration algorithm described in this section. The Providence Aerial Multiview (PAMView) dataset, produced by M. I. Restrepo at Brown University for [28] (and examined in detail in [29]), contains several videos representing 31 sites in Providence, Rhode Island containing scenery of increasing complexity. The videos were produced during a series of helicopter flights between 200 and 350 meters above ground level on a JVC JY-HD10U camcorder at a resolution of 1280×720 pixels. While only one modality is present in the dataset, the accompanying metadata gives camera calibration matrices for each frame. Frames from PAMView are shown in Figure 3.3.

3.2 Scene Complexity-Adaptive Hierarchical Registration

Registration is a crucial step in comparing fusion criteria over time, making temporal and motion analysis possible despite camera pose changes. Therefore, the first step of the proposed system is to determine a registration function for sequential pairs of images from the same stream separated in time by some specified value δ, resulting in the transformation function $\vec{x_t} = T(\vec{x_{t+\delta}})$.

Scenes varying in complexity are a regular occurrence in aerial video. Consequently, an algorithm will be developed that is capable of registering frames from aerial video in an adaptive manner, using progressively finer-resolution transformation functions as scene features such as occlusion and non-planar backgrounds are observed. The program should be capable of fast registration of flat-world scenery taken from desert or plains settings as well, while also being capable of registering frames with buildings or mountains. The algorithm should be robust to changes over time as a result of

(a) SST in EO, $t = 30$ seconds

(b) SST in IR, $t = 30$ seconds

(c) CMM in EO, $t = 7$ seconds

(d) CMM in IR, $t = 7$ seconds

(e) SMS in EO, $t = 4$ seconds

(f) SMS in IR, $t = 4$ seconds

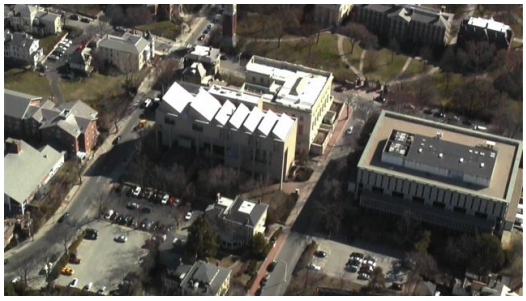
Figure 3.2: Frames from the Calamityville datset
Videos were classified with a three letter designation. The first letter, S or C, indicates smoke or clear conditions. The second, M or S, indicates motion or stationary camera. The last letter, N, S, M, or T, indicates human presence and activity: no presence, stationary targets, moving targets, or simulated threats, respectively.

(a) Site 1: An airstrip

(b) Site 2: An airstrip and hangar

(c) Site 6: A hardware store and parking lot

(d) Site 7: A suburban neighborhood



(e) Site 16: Rockefeller Library

(f) Site 23: Greene St. buildings

Figure 3.3: Frames from the PAMView dataset

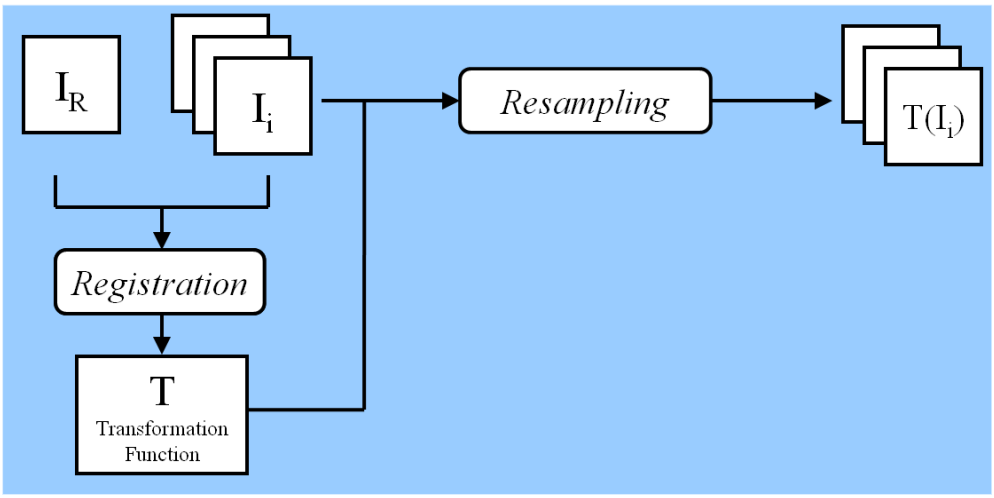


Figure 3.4: The registration subsystem

entering or exiting areas of high complexity.

Assumptions can be made about the videos used as input to the system. The frames are already coregistered in both space and time, so that the image domain of any two frames with the same index number, regardless of modality, is the same. The videos are also taken at near nadir angle, so that the horizon is not visible. Neighboring frames are also expected to have slight geometric changes at ground level. Naturally, parallax for buildings will result, but movement of major background areas is assumed to be minimal, and there is expected to be a great deal of overlap (greater than 85%) between frames.

To accomplish this goal, the following novel approach is proposed:

3.2.1 Registration Procedure

The subdivision technique is based upon the insight that, at certain scales and regions of interest, the relationships between corresponding scenes appears affine rather than projective. First, a fundamental registration procedure and image size are selected. All image analysis will take place on images of the selected size throughout the procedure, and only affine transformations will be computed. At the "top" level, the entire image is used to obtain an approximate registration of the global scene. At this level, the image is sized down by a factor of 2^n in both width and height. The subdivision technique is then applied: the image is then considered at $1/2^{n-1}$ scale with the previously-computed transformation applied via reverse resampling. At this larger scale, the image is divided into four quadrants, and the procedure repeats recursively.

As input, take I_R and I_T as reference and target images, respectively. Since frames from different modalities are coregistered, it suffices to select only one pair of images and apply the same registration across modalities. Initialize a transformation T to an identity transformation, represented as an identity matrix.

The procedure for a region is given:

1. Create a resampled estimate of the target image $I_{est} = T(I_T)$
2. Obtain landmarks L_R and L_L in I_R and I_{est} via Laplacian of Gaussian feature detector with standard deviation σ_{LoG}
3. Compute invariant moments for all landmarks
4. Determine putative correspondence by minimizing a distance metric applied to pairs of landmarks from L_L and L_R
5. Obtain inliers and affine transformation matrix A via RANSAC

If a set of subdivision conditions are satisfied, perform the following:

1. Divide the image into four quadrants
2. For each image quadrant, let T be the product of previous transformations and A, then recursively repeat the procedure.

If no subdivison takes place, the product of previous transformations and affine matrix A is returned. When all subdivision has concluded, each transformation matrix A is returned and combined via basis functions. I_T can be resampled with the final transformation, generating the aligned image $T(I_T)$

3.2.2 Landmark Detection

A Laplacian-of-Gaussian (LoG) filter is used to extract landmarks in the image. The Laplacian-of-Gaussian can be computed via convolution with the function

$$G''(\vec{x}) = -\frac{1}{\pi\sigma^2}\left(\frac{1-\|\vec{x}\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|\vec{x}\|^2}{2\sigma^2}\right) \tag{17}$$

yielding a negative response where light-colored blobs exist in the input and a positive response where dark blobs exist. Under tighter performance constraints, a Difference-of-Gaussian (DoG) filter may be substituted. Landmarks consist of local minima and maxima in the filtered image,

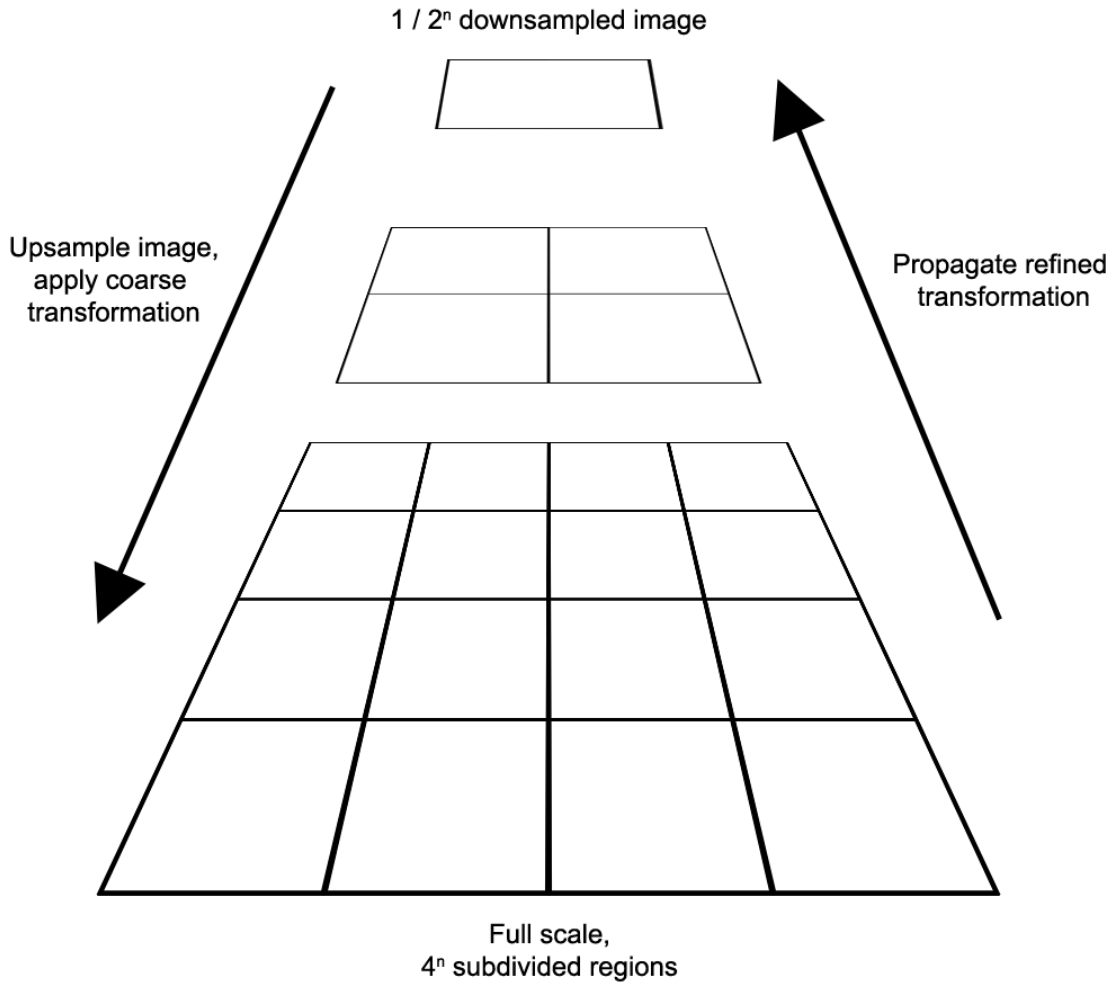
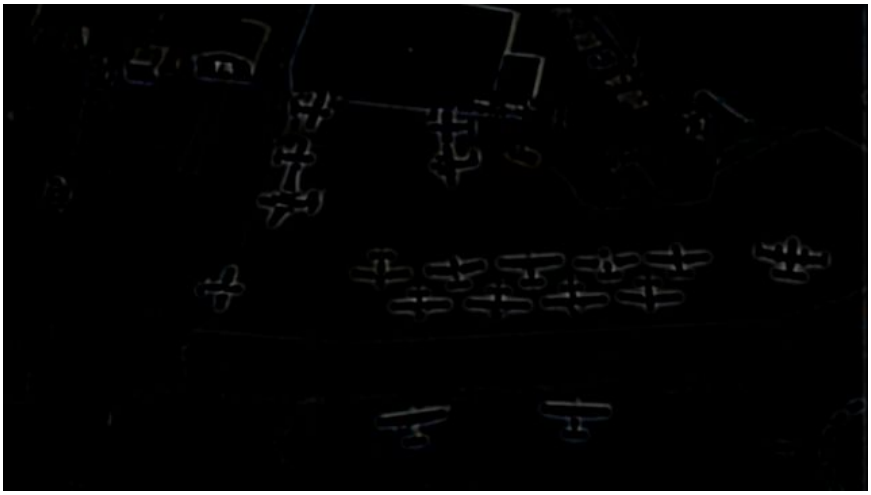


Figure 3.5: Adaptive registration procedure

(a) An image to be filtered



(b) The LoG response with $\sigma = 4.0$

Figure 3.6: The Laplacian of Gaussian filter

and extrema within a larger circular neighborhood are desirable to ensure unique, rotation-invariant landmarks.

A novel, but minor, optimization for determining local extrema is as follows: consider that each candidate location during landmark detection must be either a minimum or a maximum within a circular template of pixels centered on the candidate. Then it is possible to limit the number of comparisons, with non-extrema are eliminated as candidates as early as possible (see Figure 3.7). To accomplish this, a marking grid of identical size to the image is used to record which locations have been visited. As a candidate point is being tested for local extremity, each point within the template is compared to the candidate. If another point in the template is found to be larger (or smaller, if searching for a minimum), the candidate point is ruled out immediately, and its corresponding location within the grid is marked as visited. Consideration continues, then, at the larger value that disqualified the previous candidate, and the process repeats. If, instead, the point within the template does *not* disqualify the candidate, then that point is marked as visited, and it will not need to be considered as a candidate in the future. It is sufficient to make two passes through the filtered image, one for minima, and one for maxima. Further, it is also very simple to adjust the algorithm to consider only points higher (or lower) than a certain amount above their neighbors, as the logic remains the same, while only the comparison function changes.

3.2.3 Landmark Correspondence

Once a list of extrema is created, the locations are considered landmarks, and feature vectors for each point are computed, intended to be used to determine putative correspondence. While many schemes could be substituted to compute features, it was determined that Hu's rotation invariant moments sufficed to compute an accurate registration of regions in the datasets used [30]. A brief summary of the computational method used to obtain the features is given below.

For a circular template w centered about a landmark in image $I(x, y)$, the center of mass $(\bar{x}, \bar{y})$ is first computed as

$$\bar{x} = \frac{1}{|w|} \sum_{(x,y)\in w} xI(x,y) \quad \bar{y} = \frac{1}{|w|} \sum_{(x,y)\in w} yI(x,y) \tag{18}$$

Then the central moments M_{pq} can be given

$$\mu_{pq} = \sum_{(x,y)\in w} (x-\bar{x})^p (y-\bar{y})^q I(x,y) \tag{19}$$

Often, the moments are normalized for scale, and while this is not strictly necessary in pure rotation,

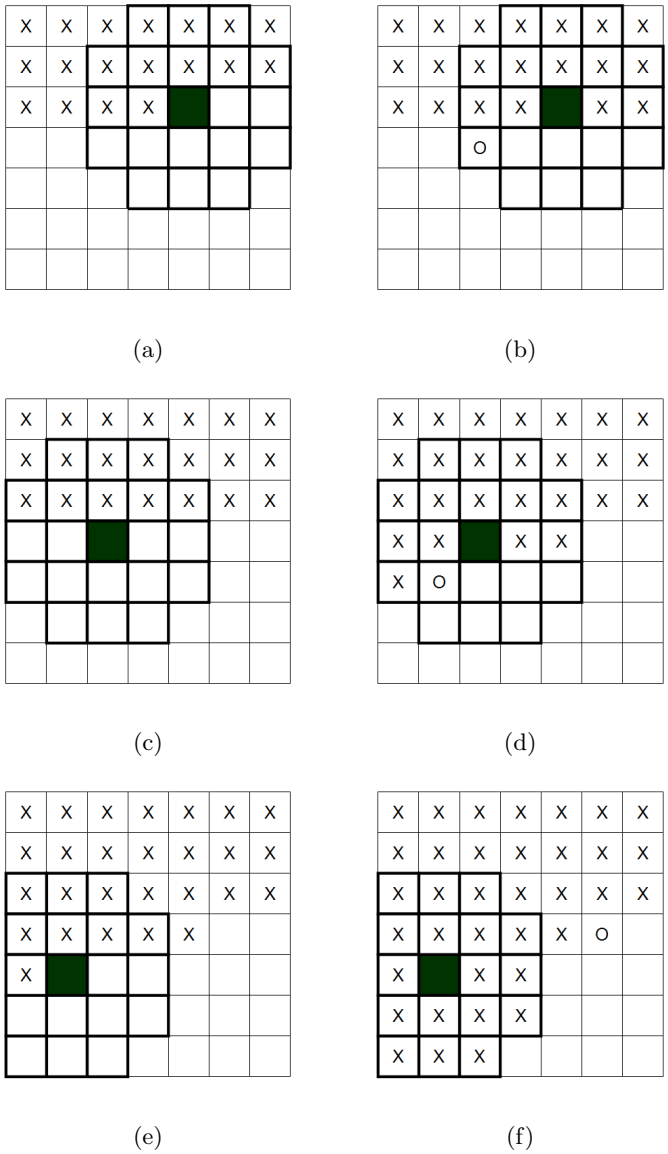


Figure 3.7: Finding local extrema
(a) An intermediate step is shown, with the neighborhood (a circular template of radius 2) drawn with heavy borders, previously-tested squares marked X, and the current candidate shaded.
(b) As the locations within the neighborhood are tested, a higher value is encountered, marked with an O in the diagram.
(c) The higher value becomes the new candidate extremum. While squares in the new neighborhood with an X must be tested again, they cannot become candidates again.
(d) Again, the test encounters a higher value, marked O in the diagram.
(e) The new candidate extremum is tested from the revised location.
(f) Finding the candidate was extreme within the neighborhood, it is accepted as a new landmark, and the square marked O will become the next candidate to be tested.

the presence of any tilt or zoom in the image may cause some scaling to occur. Normalization creates

$$\eta_{pq} = \mu_{pq} \left(\frac{2}{\text{width}(w) + 1} \right)^{p+q+2} \tag{20}$$

Finally, the seven invariant moments given by Hu are

$$\begin{aligned}
\mathcal{I}_1 = & \quad \eta_{20} + \eta_{02} \\
\mathcal{I}_2 = & \quad (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\mathcal{I}_3 = & \quad (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\mathcal{I}_4 = & \quad (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\mathcal{I}_5 = & \quad (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2) \\
& \quad +(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) \\
\mathcal{I}_6 = & \quad (\eta_{20} - \eta_0 2)((\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\mathcal{I}_7 = & \quad (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2) \\
& \quad -(\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2)
\end{aligned} \tag{21}$$

(Note: these were originally designated I_1 through I_7, but are represented here with the symbol $\mathcal{I}$ to avoid confusion with the numerous images throughout the document)

With the values of invariant moments computed at each landmark, a metric must be employed to determine correspondence. In this project, a simple Euclidean norm is used, with each moment normalized to a zero mean and a unity variance. For each landmark in the reference image, the landmark with the smallest metric distance for the features contained is chosen as a correspondence match, and recorded for the next step of the algorithm. That is, if X_R and X_T are the sets of features computed by finding extrema in the LoG-filtered reference and target images, respectively, then the correspondence is found for the first n Hu moments by:

$$\text{corr}(\vec{x_R}) = \underset{\vec{x_T} \in X_T}{\text{argmin}}\, D(\vec{x_R}, \vec{x_T}) \tag{22}$$

with

$$D(\vec{x_R}, \vec{x_T}) = \left(\sum_{j=1}^{n} N_j(\vec{x_R}) - N_j(\vec{x_T}) \right)^{\frac{1}{2}} \tag{23}$$

and N_j representing the normalized j^{th} moment,

$$N_j = \frac{\mathcal{I}_j - E[\mathcal{I}_j]}{E[\mathcal{I}_j - E[\mathcal{I}_j]]} \tag{24}$$

3.2.4 Affine Registration

Correspondences computed by the previous step are not necessarily suitable for direct use computing a transformation function, even a function as simple as an affine transformation. The presence of misclassified correspondences, or outliers, has a profound impact on the accuracy of an affine registration, and therefore, on the transformation as a whole. Further, the simplicity of minimizing the metric distance between feature vectors in the two frames being aligned allows for one-to-many/many-to-one relationships in the original correspondence set. To address these issues, determine inliers, and compute an accurate affine transformation, random sample consensus (RANSAC) will be employed [31].

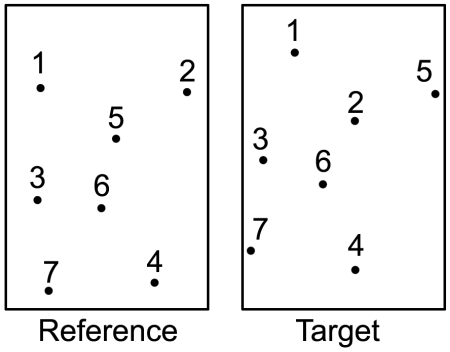
The method will take an initial set of the n putative correspondences containing outliers and attempt to select a set of inliers determining the transformation fitting the most points. The basic procedure is as follows: assume that we are attempting to find inliers among two sets of points $\mathbf{x}$ and $\mathbf{x}'$ so that $\vec{x_i} = A\vec{x_i'}$, denoted hereafter $\vec{x_i} \leftrightarrow \vec{x_i'}$.

1. Initialize a number $n_{best} = 0$ and A_{best} to an identity matrix
2. Randomly select k points as a subset of the putative correspondences
3. Compute the transformation function A defined by the selected points
4. Transform all points in $\mathbf{x}'$, creating Ax'
5. Determine n_{sample}, the number of points in $A\mathbf{x}'$ match the position of their correspondences in $\mathbf{x}$
6. If $n_{samples} > n_{best}$, set n_{best} to n_{sample} and A_{best} to A
7. If a satisfactory number of inliers is found, or a maximum number of subsets K has been examined (described below), terminate with results; otherwise, repeat from step 2.

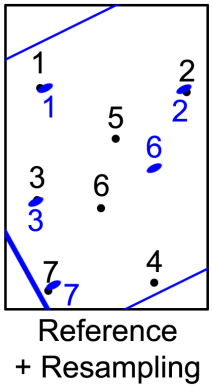
If we suppose that there are n_{good} true inliers out of the n putative correspondences matching under a correct transformation, the odds of selecting k inliers in one iteration of the method is

$$p_k = \prod_{i=0}^{k-1} \frac{n_{good} - i}{n - i} = \frac{n_{good}!}{(n_{good} - k)!} \frac{(n-k)!}{n!} = \frac{\binom{n_{good}}{k}}{\binom{n}{k}} \tag{25}$$

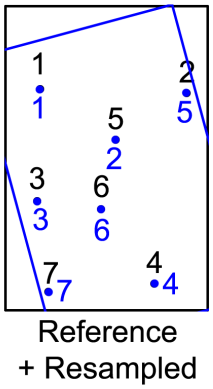
Hence, the odds of a defining a correct transformation in one pass of the algorithm is p_k.



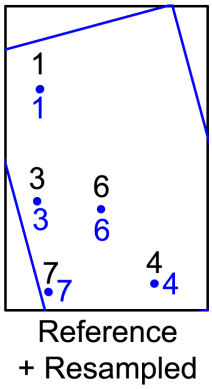
(a) Two datasets containing outliers



(b) Transformation imposed by a bad sample



(c) Transformation imposed a sample of inliers



(d) Final transformation and inliers

Figure 3.8: Unsuccessful and successful transformations during RANSAC

If K attempts are made, the odds of correctly deducing the transformation can be given as

$$P_K = p_k + (1 - p_k)P_{K-1} \tag{26}$$

with

$$P_0 = 0 \tag{27}$$

or as the summation

$$P_K = p_k \sum_{i=0}^{K-1} (1 - p_k)^i \tag{28}$$

Then, selecting a desired probability of success z, the appropriate number of iterations K can be described as

$$K = \underset{x}{\operatorname{argmin}} P_x \geq z \tag{29}$$

However, the variable n_{good} (and therefore, p_k and P_K) is rarely known, and often difficult to estimate, varying from input to input. An adaptive formula can be substituted to estimate K as the algorithm iterates, based upon the above formulation of the probabilities involved. Before the sampling begins, K is set to infinity. Then, after iteration s, K can be estimated

$$K = \frac{\log(1 - z)}{\log(1 - (1 - \epsilon)^n)} \tag{30}$$

with

$$\epsilon = \frac{1 - n_{best}}{n} \tag{31}$$

so that the subroutine yields its results when $s \geq K$.

The described approach for finding a correct transformation requires that RANSAC determine many candidate transformations of the same type as the desired result, hence, a fast solution for an affine transformation is necessary (and k must be no smaller than three).

An affine transformation is given as $\vec{x}' = A\vec{x}$, with

$$A = \begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ 0 & 0 & 1 \end{bmatrix} \tag{32}$$

and $\vec{x}$ as a homogeneous 3-vector.

Given three correspondences $\vec{x}_0 \leftrightarrow \vec{x}'_0$, $\vec{x}_1 \leftrightarrow \vec{x}'_1$, and $\vec{x}_2 \leftrightarrow \vec{x}'_2$, the affine transformation

relating them can be determined. Assuming that each homogeneous vector is normalized (i.e. $\vec{x_i} = [\ x_i \quad y_i \quad 1\ ]^T$), a system of equations is constructed to solve for the six degrees of freedom in the transformation:

$$\begin{bmatrix} x_0 & y_0 & 1 & 0 & 0 & 0 \\ x_1 & y_1 & 1 & 0 & 0 & 0 \\ x_2 & y_2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_0 & y_0 & 1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 \\ 0 & 0 & 0 & x_2 & y_2 & 1 \end{bmatrix} \begin{bmatrix} a_{00} \\ a_{01} \\ a_{02} \\ a_{10} \\ a_{11} \\ a_{12} \end{bmatrix} = \begin{bmatrix} x'_0 \\ x'_1 \\ x'_2 \\ y'_0 \\ y'_1 \\ y'_2 \end{bmatrix} \tag{33}$$

Naturally, the blockwise diagonal matrix can be broken into two systems of three equations for a simpler solution.

When RANSAC has finished, a subset of the original corresponding points will be identified as inliers, and their affine relationship will have been approximated by only three of the pairs. An optional step remains: the transformation can be recomputed using least squares to increase accuracy. Altering the above formula by multiplying both sides by the transpose of the 6×6 matrix produces the equation

$$\begin{bmatrix} \sum x_i^2 & \sum x_i y_i & \sum x_i & 0 & 0 & 0 \\ \sum x_i y_i & \sum y_i^2 & \sum y_i & 0 & 0 & 0 \\ \sum x_i & \sum y_i & n & 0 & 0 & 0 \\ 0 & 0 & 0 & \sum x_i^2 & \sum x_i y_i & \sum x_i \\ 0 & 0 & 0 & \sum x_i y_i & \sum y_i^2 & \sum y_i \\ 0 & 0 & 0 & \sum x_i & \sum y_i & n \end{bmatrix} \begin{bmatrix} a_{00} \\ a_{01} \\ a_{02} \\ a_{10} \\ a_{11} \\ a_{12} \end{bmatrix} = \begin{bmatrix} \sum x_i x'_i \\ \sum y_i x'_i \\ \sum x'_i \\ \sum x_i y'_i \\ \sum y_i y'_i \\ \sum y'_i \end{bmatrix} \tag{34}$$

Again, a blockwise solution exists, and the same matrix solution subroutine can be employed to compute a best-fit solution for the affine transformation parameters. Having computed the affine relationship, the fundamental transformations combined by the registration subsystem are now complete.

3.2.5 Subdivision Conditions

It remains to determine how many levels subdivision must take place. Consider that, for a subdivision technique such as the following to succeed, any factors that limit registration on zoomed-in parts of the images must be identified. For example: as the images are broken down, the non-overlapping

portions of the images will eventually comprise a large percentage of some pair of images to be registered. And another problem: eventually, the resampling procedure may emphasize any blurring or artifacting in the images. Finally, the further registration of subdivided areas may become a liability as previous levels will have already aligned the images to a desirable accuracy, and any new errors introduced by further attempts at registration will bring the total accuracy of the registration down.

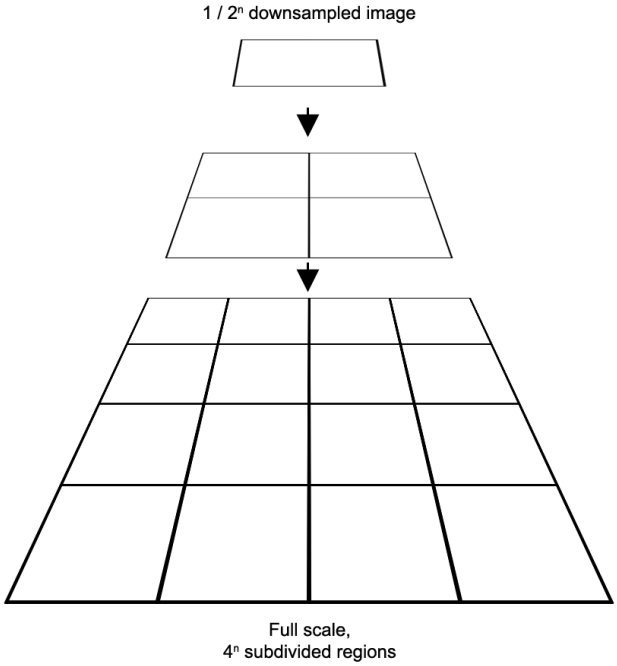
Two approaches can be formulated to address these anticipated problems, and can be considered complementary methods. One approach depends on prior knowledge and estimation of scene complexity. In this first method, a ceiling on the number of subdivisions to use is determined *a priori*. Then, as the subdivision method is computed, the fundamental registration algorithm reports the quality of the obtained registration (represented in one of a few ways, including RMSE of the two images, number of inlier correspondences, or distribution of correspondences). If the quality is below a certain threshold, or if the affine registration fails outright, the program assigns any remaining transformations to identity and ceases subdivision on the appropriate region.

The second approach can be considered the *a posteriori* counterpart to the previous method. Adjusting the algorithm, the decision of maximum subdivision level is left to runtime to decide. As the subdivision continues, a failure or insufficient-quality registration will still result in an identity and no further subdivision, but the final number of transformations is a result of the successful completions of the affine technique.

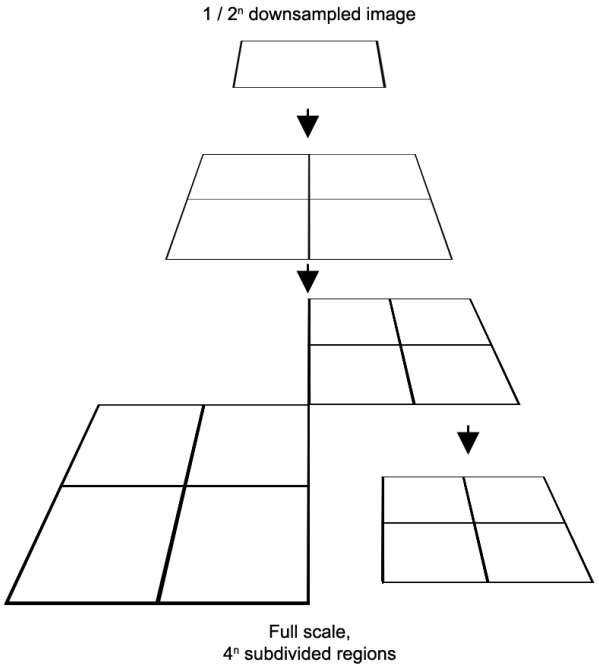
It was determined that, for the considered datasets, the *a priori* method would be preferable. A more general system may benefit from *a posteriori* methodology that adapts to a greater variety of incoming data, but for high-speed registration, the *a priori* method has two strengths: it allows for a fixed data size to be used regardless of depth, and it allows the weight function (described in the following section) to be precomputed, yielding considerable savings during final resampling and output generation. The shortcomings of the *a priori* method, and means for mitigating them, are described in Section 4.1.

3.2.6 Combining Affine Transformations

The final computation of T takes place using the results of the affine transformations described above. To reconcile the transformations obtaiend in the various areas of the image, a weighted linear approach is proposed, extending the approximation originally proposed in [32]. Each affine transformation can be represented as a matrix A_i. The matrices are added to an array structure which records the center of its affected region $\vec{v_i}$ and the standard deviation σ_i of a rational Gaussian



(a) *A priori* subdivision: fixed subimage size, predetermined maximum recursion depth



(b) *A posteriori* subdivision: subimages fixed or variable size, recursion depth decided at runtime

Figure 3.9: Subdivision rules in the registration method

basis function proportional to the size of the region.

Each affine transformation can be indexed and weighted based upon the center point $\vec{v_i}$. With a tile size given as $\vec{t}$, the centers correspond to points on a rectilinear grid, where

$$\vec{v_i} = \vec{v}_{\lfloor i/n \rfloor, i \bmod n} \tag{35}$$

so that

$$\vec{v}_{r,c} = \begin{bmatrix} c + \frac{1}{2} & 0 & 0 \\ 0 & r + \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \vec{t} \tag{36}$$

This yields a solution for the non-rigid transformation function T in terms of the n affine transformations:

$$T(\vec{x}) = \left(\sum_{i=0}^{n-1} W_i(\vec{x}) A_i \right) \vec{x} \tag{37}$$

where the weight function for the i^{th} matrix is computed as

$$W_i(\vec{x}) = \frac{G_i(\vec{x})}{\sum_{j=0}^{n-1} G_j(\vec{x})} \tag{38}$$

with Gaussian basis function

$$G_i(\vec{x}) = \exp\left\{ \frac{\|\vec{v_i} - \vec{x}\|^2}{2\sigma^2} \right\} \tag{39}$$

This novel approach is intended to replace off-the-shelf alignment methods in our use case, where camera angle and motion are expected to encounter different gradually-changing levels of complexity. In many cases, the performance of this method is expected to be sufficient to establish a consistent analysis of the scene without the overhead of a more complicated, highly localized method. The main contribution of the proposed method is the use of a subdivision approach that avoids the overhead of more complicated methods, while providing suitable accuracy to enable the use of online analysis methods.

The choice of similarity transformation results in a best-case scenario using only three degrees of freedom. If a scene is nearly planar and the camera angle is relatively stable near the nadir angle, the assumption of a single similarity transformation will be validated by this model. In that case, very little computation is required to establish T. It would be possible to choose a higher-complexity global transformation, such as an affine or projective transformation, to accomplish this goal, but

experimentation is expected to demonstrate the sufficiency of the similarity transformation.

3.3 Spatiotemporal Analysis and Decision Process

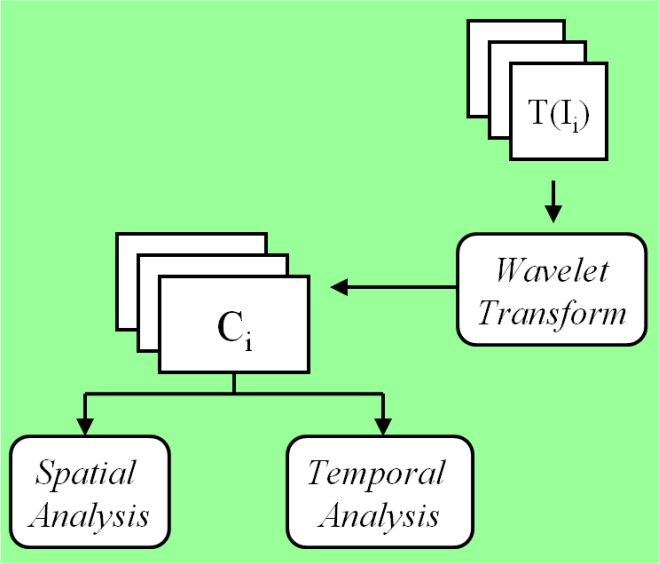


Figure 3.10: The analysis subsystem

The proposed system analyzes both regions and objects present in the video after registration has taken place to make comparisons over time possible, allowing for the decision function (and therefore, the fused result) to be based on both image and motion properties. While foreground objects are the primary concern of the proposed fusion techniques, background areas for which one modality provides significantly more structural information than the others should be considered regions of interest for fusion. As a result of the analysis step, both structural content and motion analysis should produce one unified decision map to guide the final fusion process.

Despite the seeming usefulness of background modeling methods found in literature review, a background model is not computed during analysis. Rather than classifying a pixel as foreground or background, the methods proposed attempt to compute the magnitude of motion in the frames (generally "unusual" motion, defined below), and weight the amount of motion present against the structural content in the neighborhood of each pixel. This provides a numerical value to be used by the decision function in fusion, rather than a simple assignment of class, and allows the reconciliation of spatial and temporal information in the videos.

3.3.1 Spatial Analysis in the Wavelet Domain

Of particular interest are techniques that allow selective fusion of the different sub-bands generated by the wavelet transform used in the fusion subsystem. With analysis taking place in the wavelet coefficients, instead of purely in the spatial domain at pixel resolution, the decision function can affect information from different scales of the image independently, giving an extra degree of freedom in the fusion process. Structural analysis can be expressed very simply in terms of wavelet sub-bands: the responses of the wavelet filters represent spatial information at various scales, and therefore, the magnitude of the filter responses can be used directly as a measure of structural complexity.

The construction of the spatial analysis subsystem greatly influences the formulation of the wavelet fusion: information that cannot be easily analyzed due to limitations of a certain wavelet scheme cannot easily be fused by that scheme. For that reason, the use of the discrete wavelet transform (DWT) is inappropriate: a small translation changes the coefficients greatly. Stationarity allows for greater localization of effect, despite translation of foreground objects or modification of low-frequency coefficients. The stationary wavelet transform (SWT) is a natural alternative. The SWT shares the strengths of the DWT, but allows for local modification and shift invariance. Newer schemes such as the dual-tree complex wavelet transform or the wavelet-like ridgelet and curvelet methods are less attractive, complicating the process, while adding features such as greater directional selectivity that are irrelevant to the use case at hand.

We seek to compute a set of wavelet coefficients for an image $I(\vec{x})$ to facilitate analysis. The desired coefficients will be denoted $C(\omega, \vec{x})$, where ω corresponds to the sub-band (or, equivalently, the scale). At $\omega = 0$ the highest-frequency sub-band is selected, and a total of n_ω sub-bands will be computed, with

$$n_\omega \leq \lfloor \log_2 min(\text{width}(I), \text{height(I)}) \rfloor \tag{40}$$

With the choice of the SWT, decomposition of the image can take place prior to analysis via the algorithm à trous. Recalling that the SWT is a method that extracts information in each scale via high-pass filters, while passing on progressively lower-frequency information to successive filters, the algorithm à trous gives an efficient mechanism for computing the various scales of the transform. The simplest basis that can be used with the algorithm can be initialized with the low-pass filter

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{41}$$

After the first pass of the transform, the differences between the input and the low-pass filter are extracted and stored as $C(0, \vec{x})$. Note that a family of wavelets can be determined by changing the filters used to accomplish this; a Gaussian kernel, for example, could be used to similar effect. To move on to the next sub-band, the filter must be increased in size, since the image input is not scaled (or rather, decimated) by the process, as it is with DWT. The algorithm à trous performs a filter expansion by simply separating the filter's values by zeroes, expanding the filter size by a factor of two. Hence, the second filter would be

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{42}$$

with each successive filter similarly expanded. In this way, the entire filter does not need to be applied, and the number of nonzero coefficients in the filter remains constant throughout the process. Interestingly, this also imposes a relationship of scale between the sub-bands: each represents information that is double the scale of the previous. Thus, the sub-bands could appropriately be called octaves of the image, a common term also used to describe various doubled wavelet-like filters as well as the images produced during application or inversion of both DWT and SWT.

Thus, any input image $I_i(\vec{x}, t)$ to the analysis subsystem proposed here can be decomposed to $C_i(\omega, \vec{x}, t)$ and used for spatial analysis, and the computed coefficients can be reused during motion analysis and the final fusion algorithm. Since both the orthonormal basis used for decomposition results in an overcomplete system without a unique inverse, C_i will be represented as a set of coefficients, with all operations performed independently to the elements of the set.

3.3.2 Temporal Analysis via Integral Weighted Motion

Basic motion in an image stream can be extracted via background subtraction. Treating an image stream as a spatiotemporal function simplifies the notation used: if a stream of images from a sensor is represented as $I(\vec{x}, t)$, then the change in intensity can be written

$$\frac{\partial I}{\partial t} = I(\vec{x}, t) - I(\vec{x}, t-1) \tag{43}$$

due to the discrete sampling of an image sensor, and with the variable t given in frames. Then, trivially, the summation or integration of that motion over time is also a subtraction, as

$$\int_{t_0}^{t_1} \frac{\partial I}{\partial t} \partial t = I(\vec{x}, t_1) - (I\vec{x}, t_0) \tag{44}$$

Determining how much motion is present in the video, however, is not addressed by simple subtraction of this type. We would prefer to know the magnitude of the motion, given as

$$M(\vec{x}, t) = \left| \frac{\partial I}{\partial t} \right| = |I(\vec{x}, t) - I(\vec{x}, t-1)| \tag{45}$$

Then the magnitude of motion present over an interval is the seemingly less attractive summation,

$$\int_{t_0}^{t_1} M(\vec{x}, t) \partial t = \sum_{t=t_0}^{t_1} |I(\vec{x}, t) - (I\vec{x}, t-1)| \tag{46}$$

However, a novel optimization is proposed, adapted out of the spatial domain. Integral images, used in adaptively-boosed cascades of Haar-like classifiers, were originally proposed by Viola and Jones in 2001 [33], and compute sums of rectangular areas in constant time by caching summations at each pixel location in an image. Similarly, the proposed technique stores an image-sized matrix called an accumulator image for each image in a stream to be analyzed, using a circular buffer $n+1$ frames long to analyze motion in an n-frame interval. The accumulator image $ACC(\vec{x}, t)$ is given

$$ACC(\vec{x}, t) = \sum_{\tau=0}^{t} M(\vec{x}, \tau) = M(\vec{x}, \tau) + ACC(\vec{x}, \tau - 1) \tag{47}$$

making it possible to compute integrals over many frames in constant time per pixel:

$$ACC(\vec{x}, t_1) - ACC(\vec{x}, t_0 - 1) = \sum_{\tau=0}^{t_1} M(\vec{x}, \tau) - \sum_{\tau=0}^{t_0-1} M(\vec{x}, \tau) = \sum_{\tau=t_0}^{t_1} M(\vec{x}, \tau) = \int_{t_0}^{t_1} M(\vec{x}, t) \partial t \tag{48}$$

The ability to integrate motion over various intervals allows us to finally define "unusual" motion: high-magnitude motion in a pixel location where high-magnitude motion has not been observed in recent frames is unusual, whereas high-magnitude motion in an area exhibiting high motion in recent frames is less unusual. This can be represented in terms of two integrals: one, giving the short interval of interest (called the impulse interval, or *imp*), and other other, giving the longer interval used to tell how much motion was typical in an area in recent frames (the reference interval,

ref). Thus, the unusual motion $\hat{M}$ is given by integral weighted motion (IWM) as

$$\hat{M}(\vec{x},\tau) = \frac{\int_{\tau-imp}^{\tau} M(\vec{x},\tau)\partial t + c}{\int_{\tau-ref}^{\tau} M(\vec{x},\tau)\partial t + c} \tag{49}$$

with c as a stabilizing constant to avoid both numerical degeneracy and oversensitivity to compression artifacts in the video. The integral weighted motion process is shown in Figure 3.11.

Integral weighted motion tends to emphasize the outlines of moving objects, with the variable imp used to increase the width of the outline by treating more motion as immediately relevant, and the variable ref used to control the "memory" of the system with respect to motion that has occurred in the past.

The benefits of IWM can be understood in terms of the list of weaknesses given by Toyama [16], repeated in summary here:

1. Background objects can be moved, but should be reabsorbed by a background model
2. Global illumination can change gradually
3. Global illumination can change suddenly
4. Noisy and peroidic fluctuation is possible in the background
5. Foreground objects may be incorrectly absorbed into the background
6. A large number of training frames may nto be available
7. Moving objects may be homogeneous in texture
8. Foreground objects may "sleep," becoming background objects
9. Background objects may "wake," becoming foreground objects
10. Shadows may be cast by foreground objects

By avoiding explicit classification of foreground and background objects, items 1. and 5. are immaterial. As IWM output is independent of frames older than the last ref input frames, 2., 8., and 9. have a local, limited effect. Only ref frames are required to produce an output, so 6. does not apply. Given that edges of objects are the only components expected to produce a high response, 7. does not alter the output of IWM in a meaningful way. 3. and 10. are arguably the weakest parts of IWM, but the low likelihood of sudden changes in the desired and the irrelevance of shadows as an artifact prevent them from disqualifying IWM for the problem at hand. 4., however,

(a) The impulse interval represents motion to be analyzed

(b) The reference interval collects motion over a longer interval to characterize expected motion

(c) Dividing the two yields a high response in areas exhibiting unexpected motion

Figure 3.11: Integral weighted motion

demonstrates the strongest aspect of IWM. High noise in the background creates a high response in the denominator, minimizing the importance of impulse motion in noisy areas.

3.3.3 Temporal Analysis via Wavelet-Domain Three Frame Differencing

As an alternative to integral weighted motion, finding the silhouettes of moving objects via three frame differencing may be preferable. As shown in the previous section, computing instantaneous motion at a specific frame may be done by subtracting the previous frame. If an object is moving over a span of time, subtracting two images a number of frames apart may reveal the silhouette of the object (as shown in Figure 3.12a. Unfortunately, subtracting frames in this manner creates a double image, as the images will differ at the object's previous and current locations. A third frame can be used to suppress the motion belonging to the background as shown in Figure 3.12b; the silhouetted area that is shared between the two pairs of subtracted images will be cancelled out (shown in Figure 3.12c, and a threshold can be used to remove it completely, leaving only the current silhouette of the moving object, shown in Figure 3.12d.

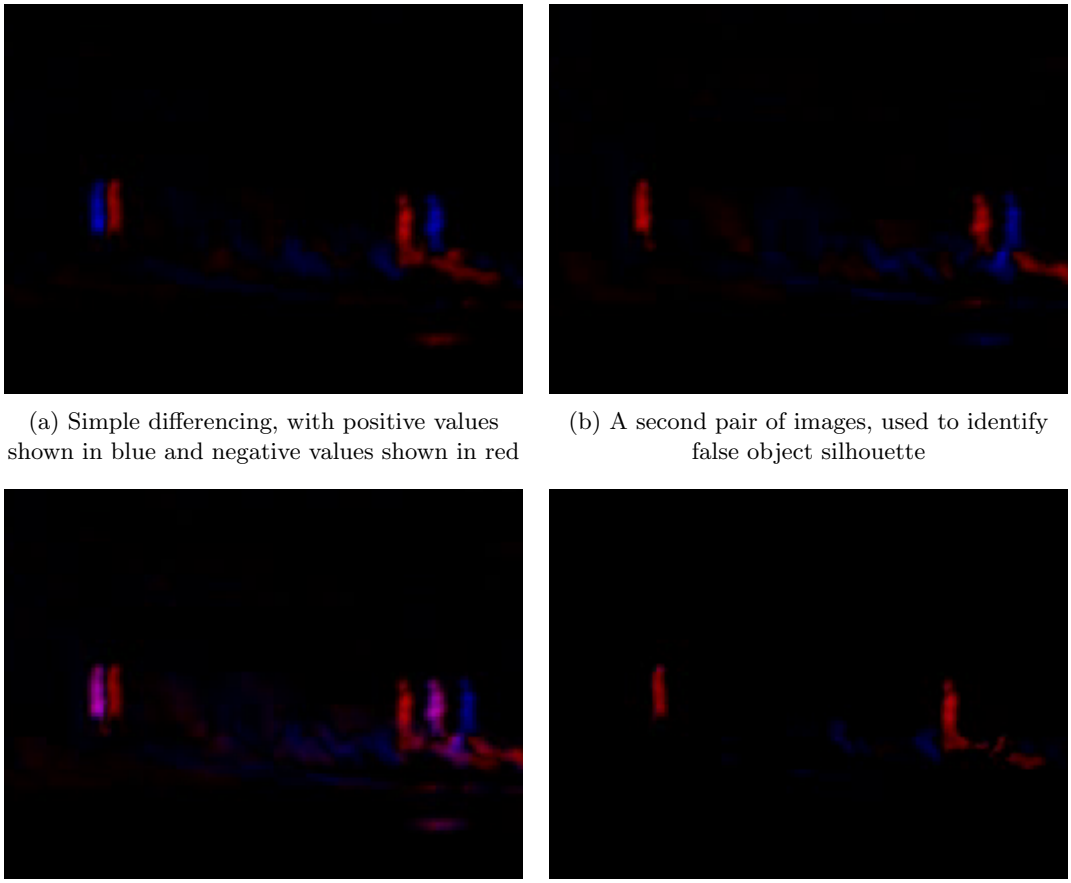
To further suppress artifacts such as changes in global illumination, or low-frequency information from semitransparent smoke, the three-frame difference can make use of the SWT coefficients already computed for spatial analysis. By selectively applying three-frame differencing to the higher-frequency sub-bands, while ignoring lower-frequency (or higher-scale) sub-bands, the differencing effectively high-pass filters the motion information, while the suppression of removed motion can take place independently on the different sub-bands. The result is an accurate silhouette that can be used even in low-resolution scenarios, or under conditions that make registration challenging.

Wavelet-domain three frame differencing can be formulated as follows: first, an interval δ is chosen, which must be sufficiently high that the objects of interest will have moved at least the width of the silhouette in the time given. Then, the formula

$$d(\omega, \vec{x}, t) = C(\omega, \vec{x}, t) - C(\omega, \vec{x}, t - \delta) \tag{50}$$

is computed to obtain a two-frame difference across each sub-band. Finally, the difference

$$D(\omega, \vec{x}, t) = d(\omega, \vec{x}, t) - d(\omega, \vec{x}, t - \delta) \tag{51}$$



(a) Simple differencing, with positive values shown in blue and negative values shown in red

(b) A second pair of images, used to identify false object silhouette

(c) The overlap of the two difference images shown in magenta

(d) The silhouette of the moving object, revealed by three frame difference

Figure 3.12: Three frame differencing for motion analysis

is computed, and the final motion response is given as

$$\hat{d}(\omega, \vec{x}, t) = \begin{cases} d(\omega, \vec{x}, t) & D(\omega, \vec{x}, t) > \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{52}$$

with motion threshold ϵ. The values of δ and ϵ must be chosen based on image and motion characteristics, and may differ from video to video.

Finally, the data structures for representing d are simple: a queue or circular buffer of $\delta + 1$ coefficient sets is sufficient to calculate d, and D does not need to be stored after its initial use. Hence, the memory requirements for d are proportional to *delta*, and the algorithm is suitable for online use.

This method can also be evaluated by the ten weaknesses given by Toyama, as in the previous section, but as the method uses only three frames at a time to measure motion, problems such

as sleeping and waking have fewer ongoing effects. A fuller silhouette is captured than in IWM, but fluctuations in the background have a greater effect. The greatest strength of this method, especially compared to IWM, is the response to global illumination change, eliminated by selecting only high-frequency sub-bands.

3.4 Local Wavelet-Domain Fusion

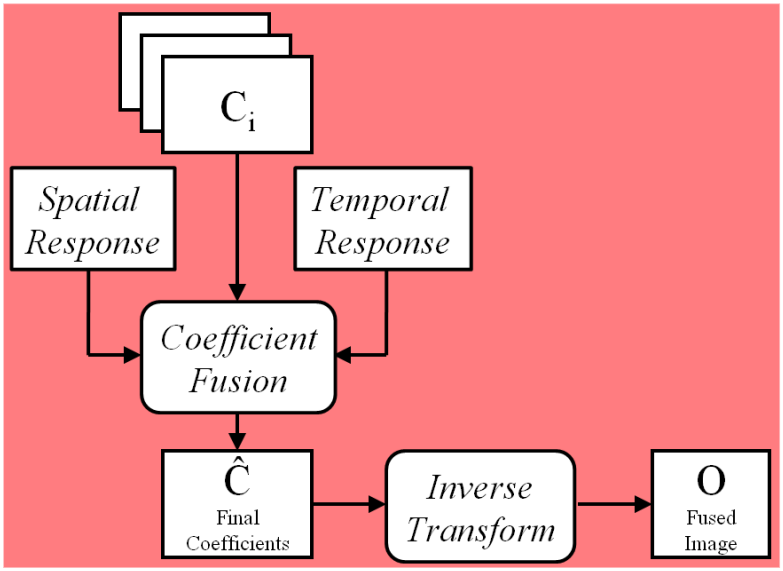


Figure 3.13: The fusion subsystem

The proposed system's final output is generated by a fusion algorithm that operates in the wavelet domain, using the comparison values generated in the analysis phase. Unlike many popular fusion methods that merge the entire image domain uniformly, the proposed method is localized, making foreground objects visible, and sensitive to both structural information and foreground motion from all modalities.

For each coefficient $C_i(\omega, \vec{x}, t)$, a comparison value α_i is generated from the results of the analysis subsystem. The vector of these values are taken as input by the decision function $DF(\alpha)$, which selects the highest value coefficient for inclusion in the fused coefficient set $\hat{C}$. Finally, $\hat{C}$ is inverted, producing an output image O.

3.4.1 The Decision Function

In frame with width w and height h, the SWT produces n_ω octaves, where $n_\omega = \lfloor \log_2 \min(w, h) \rfloor$, yielding $n \times w \times h$ coefficients per frame. For each possible octave and pixel location, the decision

function will select a coefficent modality *best*, so that $\hat{C}(\omega, \vec{x}, t) = C_{best}(\omega, \vec{x}, t)$.

The decision function is given as:

$$DF(\alpha) = \underset{i}{\text{argmax}}(\alpha_i) \tag{53}$$

with the comparison value

$$\alpha_i = k_i m_i(\vec{x}, t) + s_i(\omega, \vec{x}, t) \tag{54}$$

The comparison value is given in terms of m_i, the result of motion analysis. Depending on the methods appropriate for the input videos, m_i may be the response of IWM or the modified three frame differencing detailed above. The value s_i represents the result of spatial analysis, which in the proposed method, is exactly equal to C_i. Finally, k_i is a scaling constant used to adjust the relative importance of the motion value m_i compared to s_i. This value may differ between videos due to differences in resolution, intensity levels, and contrast characteristics.

It is possible and helpful to select only $n_f < n_\omega$ octaves to merge using the decision function, with $n_f = \lceil \log_2 h_t \rceil$ in terms of the maximum target height h_t. This reduces irrelevant low-frequency information that affects the output video without adding practical value to it. In doing so, the decision function becomes

$$DF(\alpha, \omega) = \begin{cases} \text{argmax}_i(\alpha_i) & \omega \leq n_f \\ 0 & \text{otherwise} \end{cases} \tag{55}$$

A decision function of this type no longer treats every input modality the same way, but gives preference to one modality in particular. In this way, the proposed method can be considered an asymmetric fusion technique, rather than a symmetric technique which is agnostic to the types of inputs it is receiving. The above formulation of DF gives the frame I_0 preferential treatment, and as such, it is considered the "default" modality. Selection of EO imagery, where possible, is the natural choice; the visual characteristics of visible-spectrum images will be applied, even when infrared modalities are selected by DF.

3.4.2 Composition of the Output Image

Having computed the final set of SWT coefficients $\hat{C}$, the system can produce its final output. The inversion of SWT coefficients is simply given as

$$O(\vec{x}, t) = \sum_{\omega=0}^{n_\omega - 1} \hat{C}(\omega, \vec{x}, t) \tag{56}$$

with output image O residing in the same intensity scale as the image I_0. It is possible for the fused coefficients to sum to a value outside the desired display range (usually $[0, 255]$), so an implementation of the proposed method may require the intensity to be clipped to a display interval.

As a result of the inverse SWT, the output video is now complete.

3.5 Quality Measurement

Quality measurement is not intended to be required during online operation of the proposed system. This allows for greater flexibility in evaluating the performance of the system and the usefulness of its output. While determining the "correctness" of a fused video is not a well-defined problem, the quality can be measured in a meaningful way through both manual and automated analysis.

Visual inspection of the videos should reveal concealed targets, provided the targets are visible in at least one of the input streams. A competent analyst should be able to identify regions of video containing these targets without having to repeat or slow the video's playback. A method producing equivalent results to viewing one of the unmodified input streams, or decreasing the ability of the analyst to locate moving targets, can be identified as problematic in this way.

Manual inspection without quantitative analysis has limited value. Statistical image quality measures can bridge the gap between automated quality measurement and the final product of fusion. While measures like the Wang-Bovik index [34] and quaternion-based assessment of fusion results [35] have been applied, their use is not well-suited to the problem at hand. Saliency-based methods similar to the non-reference based measure in [36] or [37] can be applied in a selective fashion to segmentation results.

Evaluating the information content of an image seems to be a desirable means of judging quality, but simply evaluating the Shannon entropy of an image or region is demonstrably insufficient. Entropy is given (in bits) as

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{57}$$

where X is a set of symbols, and $p(x)$ is the probablity of a symbol x according to the probability

density function of X [38]. This definition is very commonly used within image fusion, but has two shortcomings when applied to quality measurement.

First, if two symbols are distinct, H does not distinguish two representationally similar symbols differently from two representationally dissimilar symbols. That is, a notion of similarity does not exist at all within $H(X)$, and the representation of a string of symbols is distinct from its information in a shannon sense. When converting a range of values to a number of bins or symbols for the computation of entropy, if two values are different enough to be considered two symbols, then they are as different as possible, and if two values are similar enough to be considered one symbol, then they are as similar as two equal values. The practical result of this characteristic is that entropy cannot capture visual similarity, only order or disorder, the extent of which is related to the symbols chosen.

Second, H does not contain any representation of structure. X is given as an unordered set. If H is applied to a string, the order in which the individual symbols appear does not affect the result unless a bijection is imposed to compute H in terms of a second alphabet. In visual terms, the placement of pixel values is irrelevant to their entropy, despite the intuitive definition of entropy as a model of order or disorder. If two images are created with 50% of their pixels black and 50% white, they will have the same entropy, even if one image has been created by random placement of white and black pixels and the other has been created with all light pixels placed in the upper half of the image.

Given these two shortcomings, the original goal for which entropy might have been chosen seems difficult to reach. If the right scheme is chosen, however, both can be overcome. To represent similarity between color and intensity characteristics, incorporating simple statistical measures such as mean and variance can improve the measure. To represent placement, correlation between images or subdividing the image and computing each window's characteristics can be employed. A convenient measure incorporating a scheme such as described was given by Piella, and was selected as appropriate for determining fusion quality in the proposed system [36]. An outline of the method is given as follows:

Wang and Bovik formulated a quality measure for comparing the structural similarity (SSIM) of two images [34]. Their measure, unlike methods use as RMSE or signal-to-noise ratio, attempted to capture similarity between two images a and b in three terms: correlation, given as

$$\frac{\sigma_{ab}}{\sigma_a \sigma_b} \tag{58}$$

mean luminance,

$$\frac{2\mu_a\mu_b}{\mu_a^2+\mu_b^2} \tag{59}$$

and contrast,

$$\frac{2\sigma_a\sigma_b}{\sigma_a^2+\sigma_b^2} \tag{60}$$

with the final product (after cancellation of terms and addition of stabilizing constants c_1 and c_2) given as:

$$Q_0(a,b) = \frac{2\mu_a\mu_b + c_1}{\mu_a^2+\mu_b^2+c_1}\frac{2\sigma_{ab}+c_2}{\sigma_a^2+\sigma_b^2+c_2} \tag{61}$$

The primary difficulty in using Q_0 to capture the quality of fusion is the lack of a reference image to use as a gold standard. Piella's formula makes use of Q_0, with the fused image f compared to a and b depending on the relative saliency of each within a given window of comparison.

First, a saliency measure $s(I)$ is selected. Then, the relative saliency of an image a within a window w can be given

$$\lambda_a(w) = \frac{s(a|w)}{s(a|w)+s(b|w)} \tag{62}$$

noting that

$$\lambda_b(w) = 1 - \lambda_a(w) \tag{63}$$

By splitting the region of interest for quality measurement into a set of windows W within the region, the fusion quality can then be written as the summation

$$Q(a,b,f) = \frac{1}{W}\sum_{w\in W}\left(\lambda_a(w)Q_0(a,f|w) + \lambda_b(w)Q_0(b,f|w)\right) \tag{64}$$

Finally, Piella gives a perceptually-weighted version of Q as a slight modification of the above formula. It is this perceptually-weighted form that is used to evaluate the proposed method. A perceptual weight function is given as

$$C(w) = \frac{s(a|w)+s(b|w)}{\sum_{w'\in W} s(a|w')+s(b|w')} \tag{65}$$

Then Q is rewritten

$$Q(a,b,f) = \sum_{w\in W} C(w)\left(\lambda_a(w)Q_0(a,f|w) + \lambda_b(w)Q_0(b,f|w)\right) \tag{66}$$

Piella's formulas can be extended to n images in the ordered set $\mathbf{I}$. Redefining the weights as

$$\lambda_i(w) = \frac{s(I_i|w)}{\sum_{j=0}^{n-1} s(I_j|w)} \tag{67}$$

and

$$c(w) = \frac{\sum_{i=0}^{n-1} s(I_i|w)}{\sum_{w \in W} \sum i = 0^{n-1} s(I_i|w)} \tag{68}$$

Q can be rewritten as the summation

$$Q(\mathbf{I}, f) = \sum_{w \in W} c(w) \left(\sum_{i=0}^{n-1} \lambda_i(w) Q_0(I_i, f|w) \right) \tag{69}$$

The above formula, originally invented in single-modality fusion, is the assumption of a symmetric fusion method. That is, if λ_a is high, the image quality will be maximized only if the corresponding window resembles image a in mean intensity, contrast, and covariance. However, in multimodal fusion, representation and information are distinct. Asymmetric fusion methods can be thought of as an augmentation of one specific modality; information from other input modalities is transformed to resemble the primary modality. In the problem stated, the color EO input stream's display properties are highly desirable, and information originating in the IR band is best unified with the EO if there is no large difference in structural similarity.

Therefore, a novel fusion quality measure is proposed to measure the output of asymmetric fusion. As in Piella's formula, SSIM will be used to measure similarity to the primary modality. What remains is an appropriate measure for capturing the information selected from the secondary modalities, such as the IR band. Information theoretic measures, which were inappropriate for representation-sensitive measurement of image properties, can now be used to capture the properties in question.

Mutual information is given as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{70}$$

representing the information shared by the two random variables X and Y, with $I(X;Y) = 0$ for independent variables. It can be written in terms of the joint entropy, $H(X,Y)$ as

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{71}$$

with

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x,y) \tag{72}$$

Note that $H(X,X)$ reduces to $H(X)$ in the above equation, since $p(x,x) = p(x)$. Consequently, $I(X;X) = H(X)$.

Unfortunately, mutual information on its own produces a result in bits, rather than the desired intervals of $[-1,1]$ or $[0,1]$. Consider a use case where a is the primary modality for fusion with b as the secondary modality (playing a role similar to IR images in the collected data), and the fused results are contained in image f, as in Piella's formula. Then the ratio

$$\frac{I(b;f)}{H(b)} \tag{73}$$

represents the mutual information between b and f compared to the total information in b. If symbols in b and f are independent – that is, if f captures none of the information in b – this ratio reduces to zero. If f captures all of the information in b, then the ratio reduces to 1. Thus, the mutual information is normalized to the appropriate range for the quality measure.

While this would seem to be an adequate measure for measuring the secondary modalities, the given ratio has a significant weakness: if a and b contain some of the same information, a quality measure already incorporating structural similarity between a and f will biased heavily towards a. A measure of the information in b not present in a, but still represented in f, is highly desirable.

The denominator of such a ratio is easy to describe: information present in b, but not shared by a is defined as $H(b|a)$, the conditional entropy of b given a, computed

$$H(Y|X) = H(Y) - I(X;Y) = H(X,Y) - H(X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)}\right) \tag{74}$$

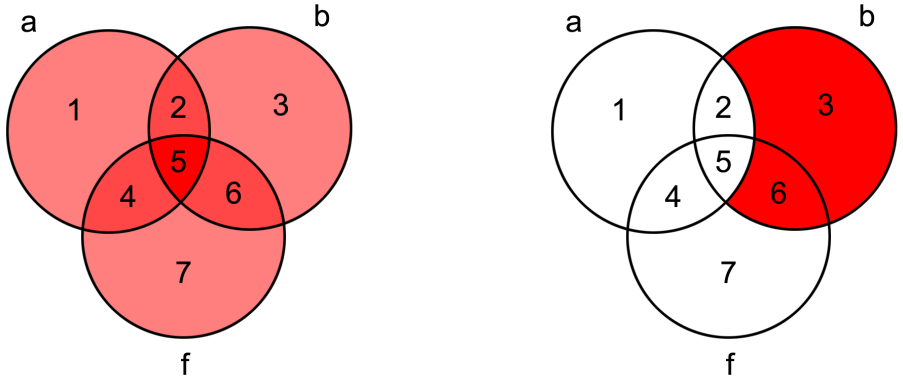
The numerator, however, is somewhat more complicated: the information shared in b and f, but not shared by a, is desired. Manipulating the joint entropies of all three variables (see Figure 3.14), the desired value can be obtained, so that the final measure of unique information shared by b and f is given

$$Q_1(a,b,f) = \frac{H(a,b) + H(a,f) - H(a,b,f) - H(a)}{H(b|a)} \tag{75}$$

With Q_0 measuring structural (or rather, representational) similarity, and Q_1 measuring the

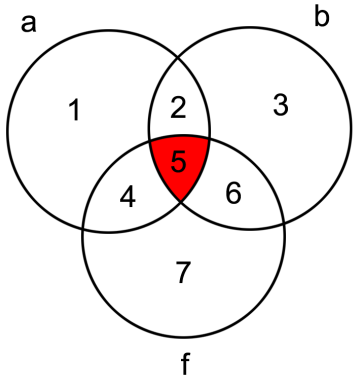
amount of unique information selected for fusion, a final asymmetric fusion measure is formulated:

$$Q_A = \sum_{w \in W} c(w) \frac{Q_0(a, f|w) + Q_1(a, b, f|w)}{2} \tag{76}$$

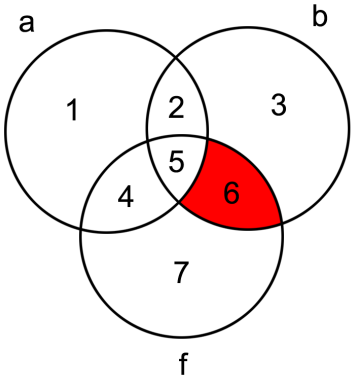


(a)

(b)



(c)



(d)

Figure 3.14: Information shared by images b and f, but not a
(a) Treating the three images as random variables, information can be shared by all three (region 5), by two variables (regions 2, 4, and 6), or only one (regions 1, 3, and 7)
(b) The conditional entropy $H(b|a)$ is relevant to the quality measure, but an equation for the portion shared by f (region 6) must be found
(c) The information shared between all three images can be found as $H(a) + H(b) + H(f) - H(a,b) - H(a,f) - H(b,f) + H(a,b,f)$
(d) Subtracting this shared information from $I(b;f)$ yields the numerator for Q_1

4 Results

4.1 Registration

4.1.1 Single-Modality Results

As described in Section 3.2.5, two approaches can be employed for adaptive registration. The *a priori* subdivision method is chosen for practical reasons: since the final number of subimages is decided upon from the beginning, the upscale-and-subdivide step of registration results in the same size image in every step of registration, and the weighted basis function is consistent among frames, lending it to precomputation and optimization (whereas deciding subdivision parameters during runtime results in unpredictable tile sizes and basis function coefficients). Therefore, the optimum tile size and maximum subdivision depth are immediately connected, and must be decided prior to registration.

Registration was applied to frames of the PAMView dataset at a tile size of 320×240, allowing two levels of dyadic subdivision. Intuitively, the use of two subdivisions allows for 16 different sets of affine parameters, sufficient for the majority of scene conditions present in PAMView: if fewer affines were used, it would be difficult to approximate projection, if a greater number were used, many regions would contain very few landmarks.

Nevertheless, many frames contained areas of high texture and areas of low texture, with subdivision only making the sparsity of certain regions more acutely felt (see Figure 4.1). The method would need to allow for sparse regions to have a sane result, even if determining a distinct affine is impossible. Additionally, some frames may be extremely challenging for landmark correspondence, even with RANSAC (see Figure 4.2). Logic for handling both misregistration and sparsity is crucial to ensure consistent results in real-world imagery.

Establishing reliable putative correspondence regardless for regions containing various levels of complexity is crucial, even with RANSAC's removal of outliers. If the number of inliers is small compared to the number of outliers, RANSAC will fail to find the correct transformation in a reasonable number of iterations. To eliminate low-quality correspondence and speed up the RANSAC process, a simple mechanism was employed: putative correspondences were established by minimizing a metric on the feature vectors. That is, if $\vec{x_T} = \text{corr}(\vec{x_R})$, all corresponding pairs with

$$D(\vec{x_R}, \vec{x_T}) > \tau_D \tag{77}$$

are eliminated prior to RANSAC. With the first five Hu moments used to define correspondences,

(a) A frame from PAMView site 2

(b) A region rich in features

(c) A region with sparse features

Figure 4.1: Feature-rich and sparse areas in PAMView

(a) A region containing redundant features

(b) A high number of features, with difficulty establishing putative correspondence

Figure 4.2: Challenging regions for landmark correspondence

$\tau_D = 3.0$ was a suitable value to suppress very poor correspondences.

A "bailout" mechanism was also implemented to handle sparse or low-confidence results from RANSAC (see Figure 4.3). If n_{best} inliers are selected from n putative correspondences by RANSAC's final iteration, n_{best} must satisfy two conditions:

$$n_{best} \geq n_{min} \tag{78}$$

and

$$\frac{n_{best}}{n} \geq \tau_n \tag{79}$$

The first condition guarantees that under sparse conditions, a minimum number of correspondences are required to establish an affine transformation. Experimentally, an n_{min} between five and ten points is sufficient to prevent sparse feature sets from establishing spurious local transformations. The second condition guarantees that even with a greater number of correspondences, a poor RANSAC result does not poison the transformation. τ_n can be set to a low percentage, with $\tau_n = 0.2$ sufficient in the majority of cases. If one of the two conditions is not met, an identity transformation is returned instead of the result computed by RANSAC, and further subdivision results are replaced by identity matrices. Note that this does *not* mean that the bailed-out regions of the image are defined by an identity transformation in the final result. Rather, the regions do not provide any *additional* adjustment to the higher-level or global transformation functions.

Blostein et al. [39] give the relationship between an LoG filter's standard deviation parameter and the size of landmarks yielding the highest response to the filter as

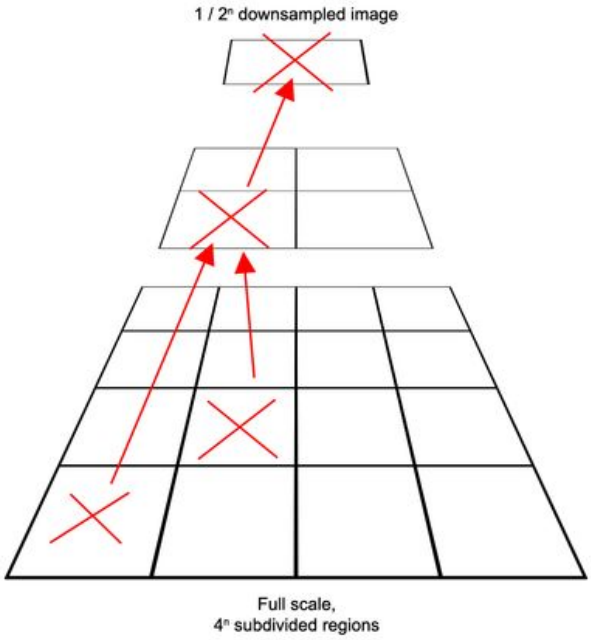
$$\sigma = \frac{r}{2\sqrt{2}} \tag{80}$$

therefore the value of σ can be reasonably chosen as 4 pixels. The value of σ with respect to the radial basis function should be proportional to the tile size; a value of $\sigma = \text{width}/2$ was selected across all videos. Finally, a value of 16 pixels was selected as the minimum distance between landmarks, with the strongest landmarks in the region given precedence.

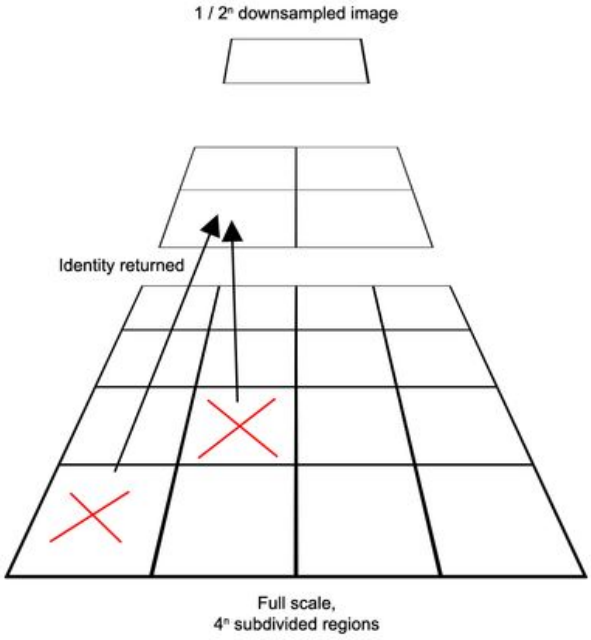
Results of registration on different videos are shown in Figure 4.4.

4.1.2 Synchronization and Coregistration

The Calamityville dataset used several computer-controlled cameras in synchronization, but the manually-controlled IR camera required the manipulation of buttons on the device, introducing a

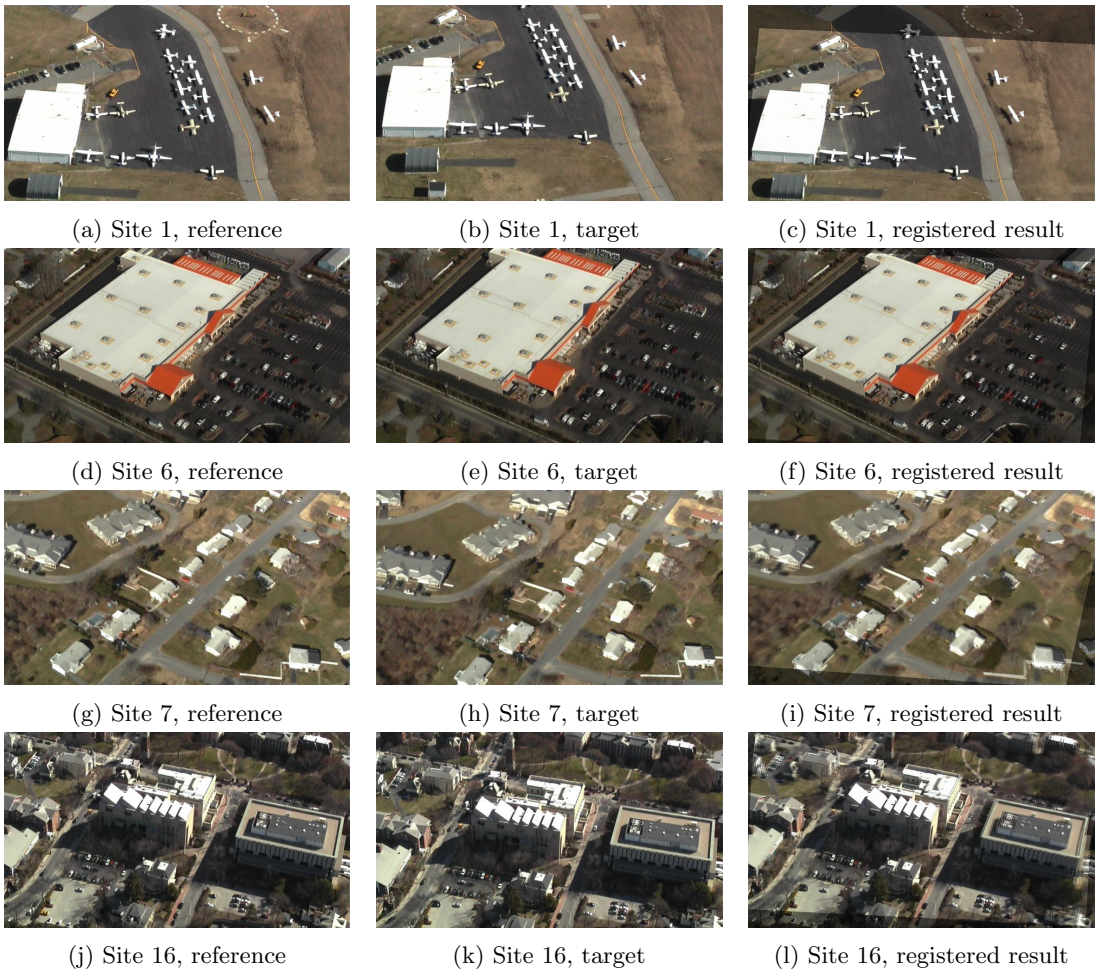


(a) A fully defined registration function containing errors



(b) A registration function employing bailout

Figure 4.3: Bailout mechanism in the adaptive subdivision registration method
In some frames, an affine registration function cannot be meaningfully established for some scales and regions. The incorrect results of local registration poison the global transformation, even if the majority of transformations are correct. Therefore, the bailout technique halts subdivision in poorly-defined areas, with all remaining levels replaced by identity matrices.



(a) Site 1, reference (b) Site 1, target (c) Site 1, registered result

(d) Site 6, reference (e) Site 6, target (f) Site 6, registered result

(g) Site 7, reference (h) Site 7, target (i) Site 7, registered result

(j) Site 16, reference (k) Site 16, target (l) Site 16, registered result

Figure 4.4: Results of the adaptive subdivision registration method

number of problems: the begin and end times for video capture would not accurately coincide across videos of each scenario, and the mechanical action of starting and stopping the video would cause misalignment in frames near the video endpoints. Since the rate of frame capture was equal between cameras, a coregistration tool was devised to correctly align and synchronize the multimodal videos, making the assumption that the majority of frames between the jitter introduced at the beginning and end would have a fixed relationship.

The tool allows for interactive resizing and shifting of one modality, while displaying both videos simultaneously (see Figure 4.5). The reference video, shown in red, is fixed, while the target video being transformed, shown in cyan, is warped to match the reference. Different keys allows for one or both videos to advance in real time, allowing different frames to be shown in each. Their offset, in frames, is subtracted from the target video if the target video's initial frame falls before the reference video's initial frame; if the reference video begins first, black frames are introduced into the target. Both nearest-neighbor and bilinear interpolation are available for scaling. Finally, when the videos are aligned, the overlapping intensities will appear gray or black, and the red or cyan tint will be minimized.

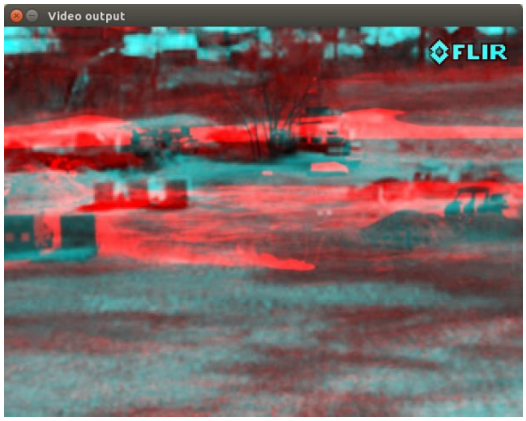
The selection of the video to transform depends on the use case. In many applications, the high-resolution data should be transformed, and will degrade less under transformation. In this case, transforming the IR is more appropriate, since the resolution and visible area of the IR would be undesirable in the result video. Nevertheless, this alignment reduces the difficulty of registering the data; if the same modality is chosen as the reference when aligning videos as is used for registration over time, the transformation obtained can be propagated gracefully across modalities, and analysis can begin.

4.2 Analysis

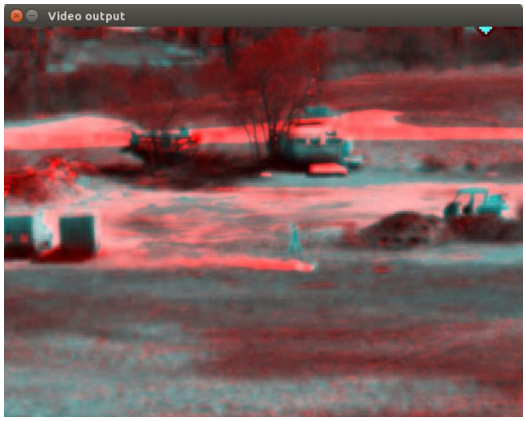
4.2.1 Preprocessing

To simplify the tasks of both analysis and fusion, preprocessing was applied to the videos. In particular, the contrast and intensity characteristics of the infrared video were transformed via SWT coefficient modification according the the formula

$$C'(\omega, \vec{x}, t) = kC(\omega, \vec{x}, t) \tag{81}$$



(a) The initial frames of two videos, prior to alignment



(b) Correct alignment of the two videos, obtained interactively

(c) The original appearance of the target video

(d) The post-alignment appearence of the target

Figure 4.5: Interactive coregistration of EO and IR videos

This has two practical results: it simplifies spatial analysis, as the low contrast characteristics of the IR bias the fusion algorithm towards the EO stream; it also inverts the coefficient bands related to the foreground objects of interest in the video so that the sign of the coefficients for the objects match across modalities, allowing consistency of representation.

The global intensity can also be modified by altering the lowest frequency sub-band. With n_ω coefficient bands, the global intensity is altered by setting

$$C'(n_\omega - 1, \vec{x}, t) = v \tag{82}$$

thus minimizing any clipping that might occur if the higher-contrast coefficients' sum approaches the boundaries of the display range.

For the Calamityville dataset, the IR band was modified with $k = -1.5$, and $v = 128$, giving an appropriate visual effect. The results of preprocessing in the IR band are shown in Figure 4.6.

4.2.2 Motion Analysis

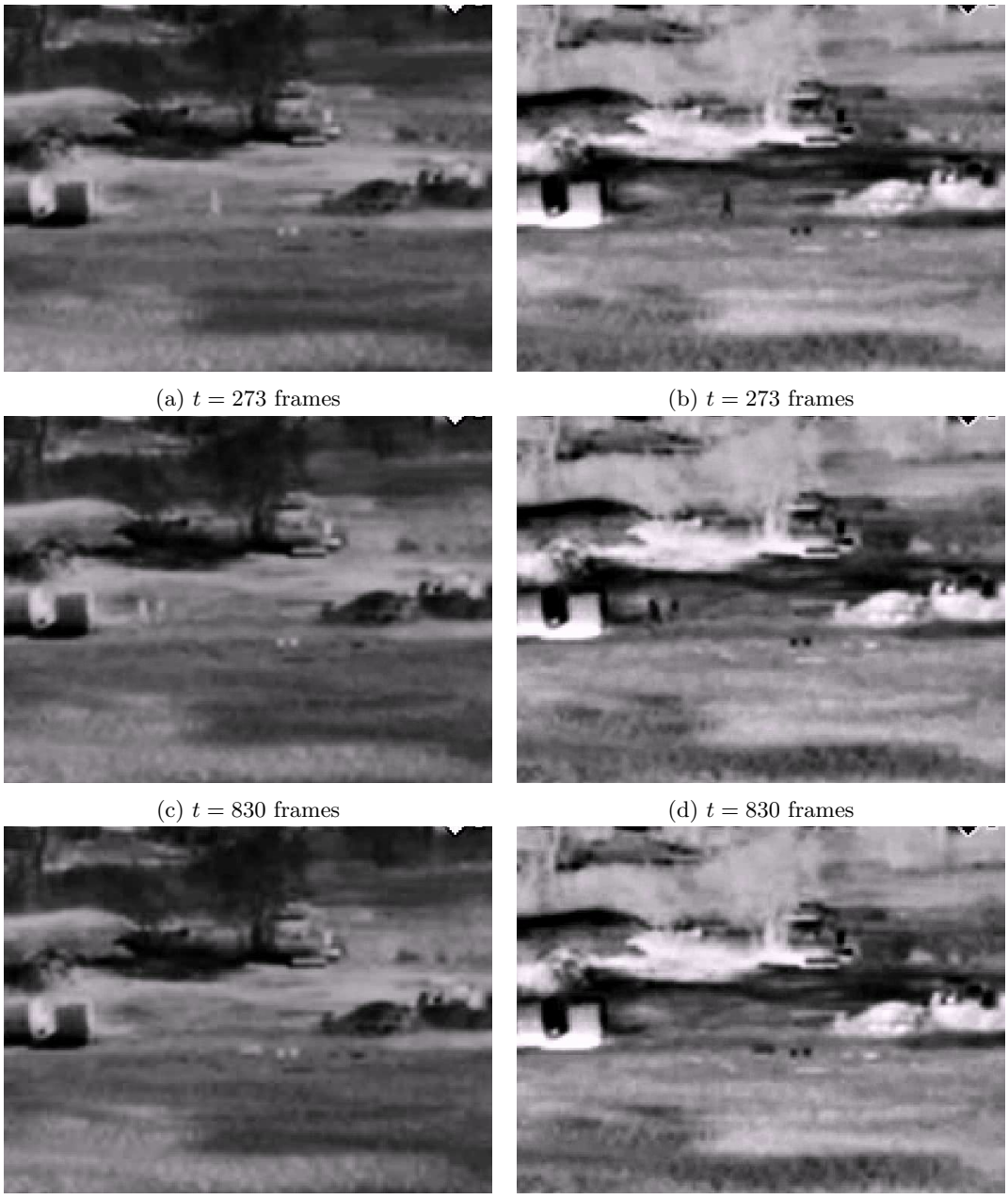
Motion in the Calamityville dataset was computed with both integral weighted motion and wavelet-domain three frame differencing in order to select the more appropriate of the two for the decision function. With IWM, selecting appropriate values for ref and imp, the two time intervals of interest. For imp, a small value is desirable, but if the value is too small, the outline of the object will be incomplete and noisy. For this reason, an imp of two frames was chosen (see Figure 4.7).

Selecting ref is much simpler. As the value of ref increases, the amount of time motion is remembered increases, and therefore, impulse motion is less likely to be unusual. Values on the order of half a second to two seconds are appropriate, with $ref = 15$ frames selected for the given datasets.

Motion captured by the IWM measure was observed in the Calamityville dataset. In the electro-optical video, a crisp outline was easily obtained. Unfortunately, the infrared video's resolution was sufficiently low, and the noise was sufficiently high, that a recognizable outline was impossible to extract via IWM (see Figure 4.8).

Under different circumstances, the performance of IWM might have been superior for multi-modal comparison, as evidenced by the sharp, consistent outline produced in EO. The alternative, wavelet-domain three-frame differencing, produced a more reliable visual result in the Calamityville dataset, while the numerical results were very close to the IWM.

Selecting parameters is sufficiently easy for the dataset. In both videos, the targets of interest

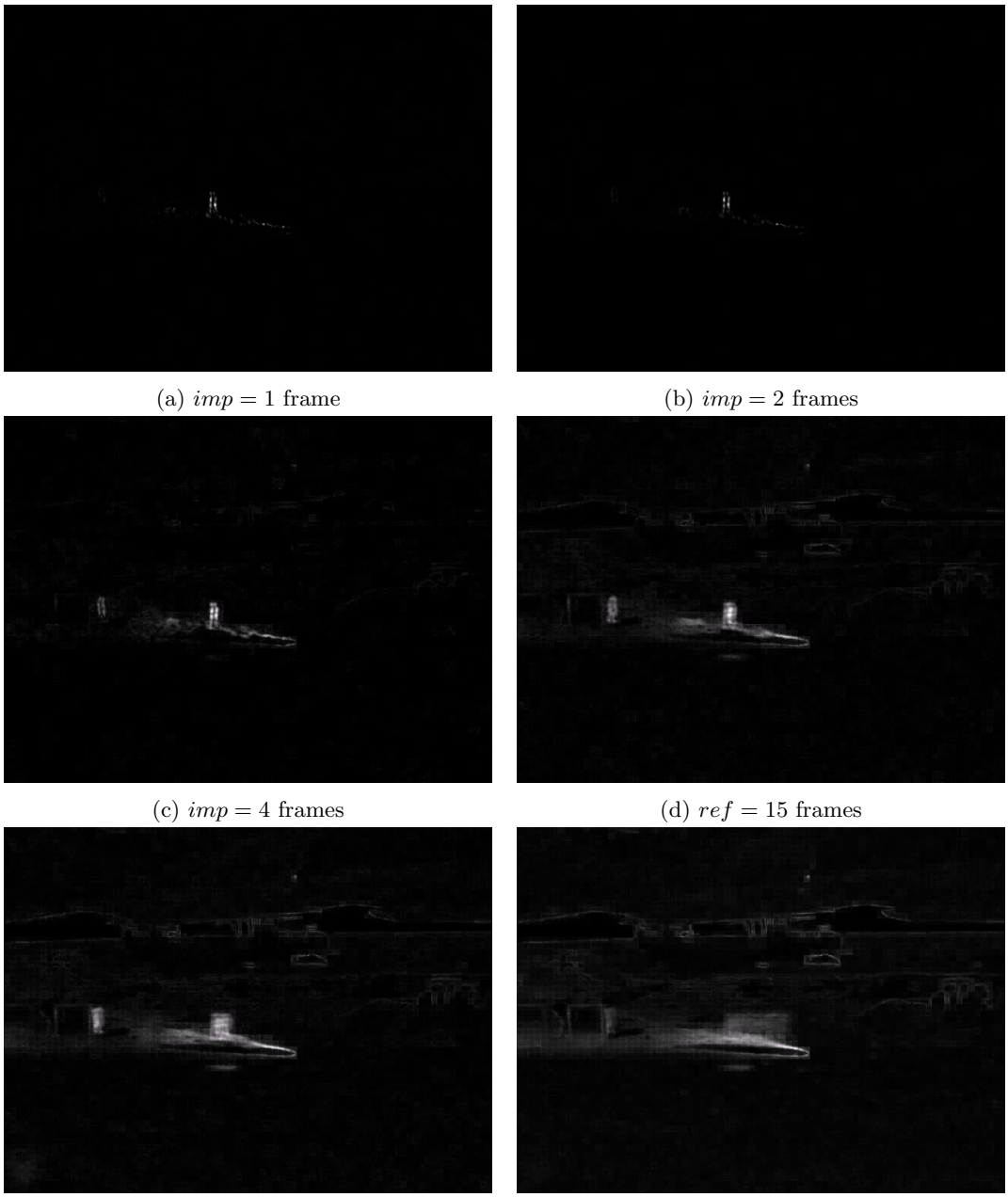


(a) $t = 273$ frames (b) $t = 273$ frames

(c) $t = 830$ frames (d) $t = 830$ frames

(e) $t = 945$ frames (f) $t = 945$ frames

Figure 4.6: Preprocessing in the IR band
Original frames shown on the left, results of contrast adjustment on the right.



(a) $imp = 1$ frame

(b) $imp = 2$ frames

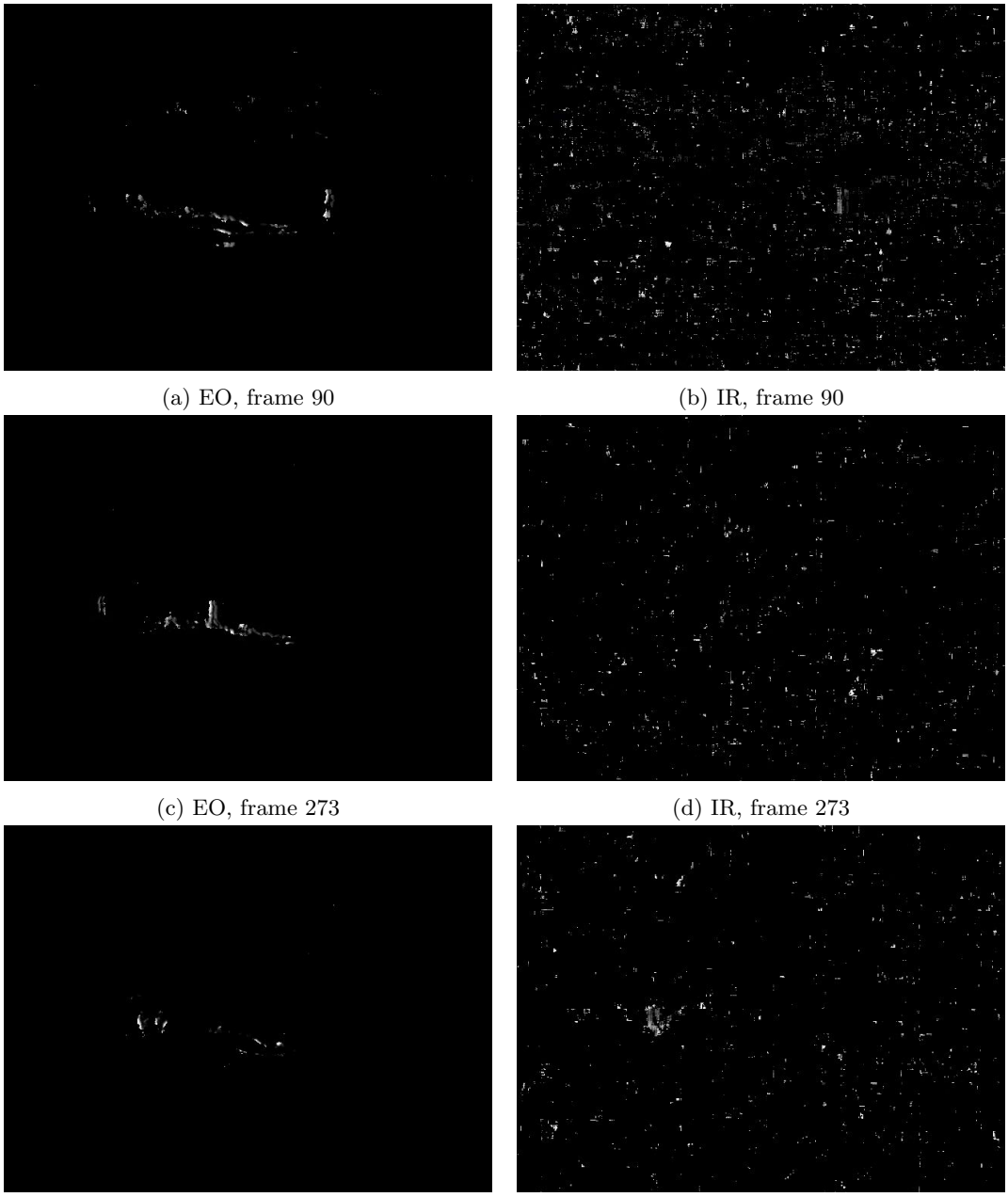
(c) $imp = 4$ frames

(d) $ref = 15$ frames

(e) $ref = 30$ frames

(f) $ref = 100$ frames

Figure 4.7: Various intervals in IWM
Note: the designation as imp or ref denotes the role such an integration would likely play, but the same integration method is used in both cases.



(a) EO, frame 90

(b) IR, frame 90

(c) EO, frame 273

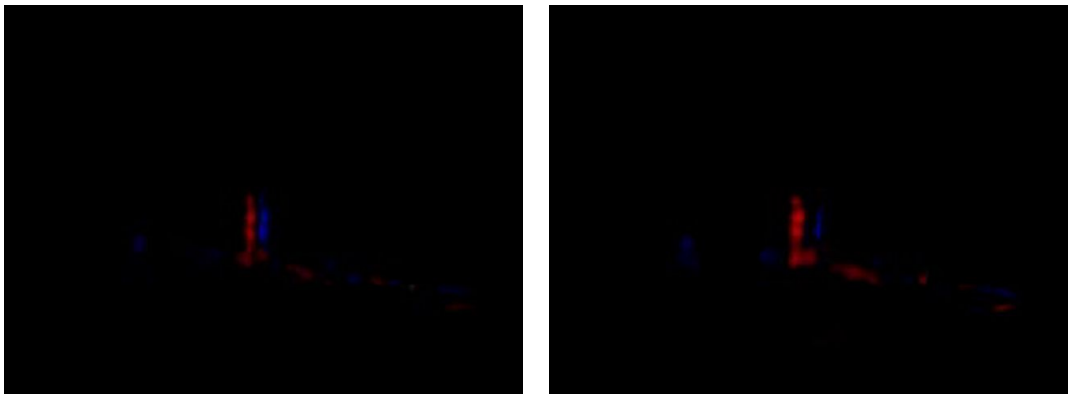
(d) IR, frame 273

(e) EO, frame 830

(f) IR, frame 830

Figure 4.8: Comparison of motion captured by IWM in two modalities

move at a consistent speed, and δ, the motion interval, is established at 20 frames (see Figure 4.9). The motion threshold, ϵ can be set to a low value, with a higher value only useful to suppress lower-amplitude motion in each scale, and was chosen arbitrarily at a value of 30.



(a) $\delta = 5$ frames (b) $\delta = 10$ frames

(c) $\delta = 20$ frames

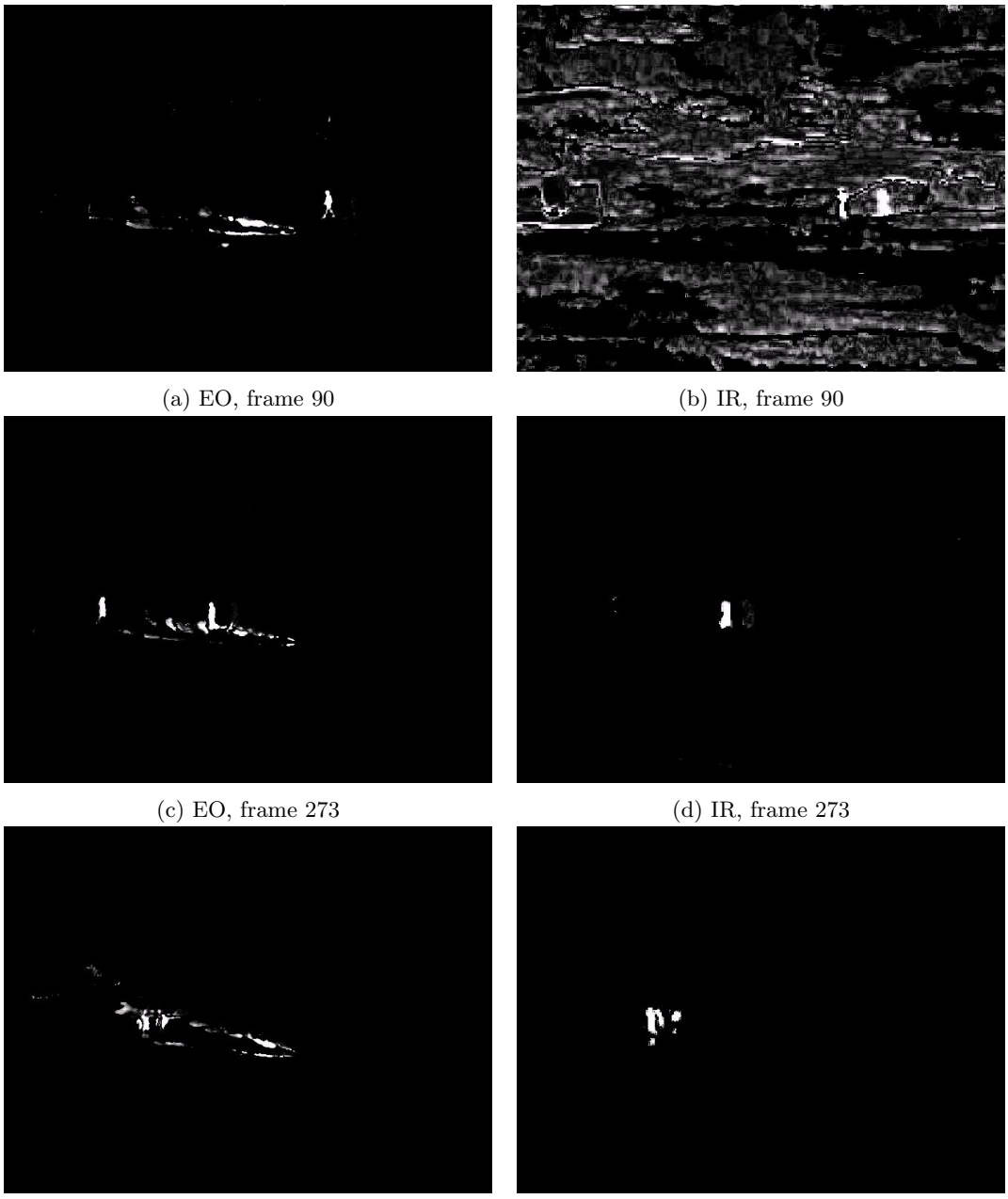
Figure 4.9: Selection of the variable δ for motion analysis
Frame 273 from the EO modality is shown above.

Results for wavelet-domain three frame differencing are shown in Figure 4.10. The selected parameters produce a good silhouette in the EO modality. While challenging, the IR dataset receives an adequate silhouette. Interestingly, the variables δ and ϵ can be set to the same value for both modalities, since the videos are at the same scale (thus, the objects are moving at the same speed) and the contrast is sufficiently similar.

4.3 Fusion

4.3.1 Colorspace Transform

The IR imagery available, unlike the EO, is a single channel. While it is possible to either decrease the EO to grayscale or treat the IR as three identical channels for the purposes of fusion, a more



(a) EO, frame 90

(b) IR, frame 90

(c) EO, frame 273

(d) IR, frame 273

(e) EO, frame 830

(f) IR, frame 830

Figure 4.10: Comparison of motion captured by wavelet-domain three frame differencing Note that the IR displays extreme sensitivity to jitter in frame 90.

desirable approach is to transform the RGB color values of the EO stream into the hue-saturation-luminance (HSL) colorspace (see Figure 4.11) and impose the fusion results upon the saturation and hue present in the EO, a process similar to the IHS method for pan sharpening [40].

Given values r, g, and b at a pixel in the interval $[0, 1]$, the transform is as follows: first, the hue h (in degrees) is found by

$$h = 60 \times \begin{cases} 0 & c = 0 \\ \frac{g-b}{c} \mod 6 & \max(r,g,b) = r \\ \frac{b-r}{c} - 2 & \max(r,g,b) = g \\ \frac{r-g}{c} + 4 & \max(r,g,b) = b \end{cases} \tag{83}$$

with $c = \max(r, g, b) - \min(r, g, b)$. Then, the lightness l can be given

$$l = \frac{\max(r,g,b) + min(r,g,b)}{2} \tag{84}$$

and the saturation s is given

$$s = \frac{c}{1 - |2l - 1|} \tag{85}$$

Transforming back to RGB after the lightness value has been altered can be accomplished with the following: With

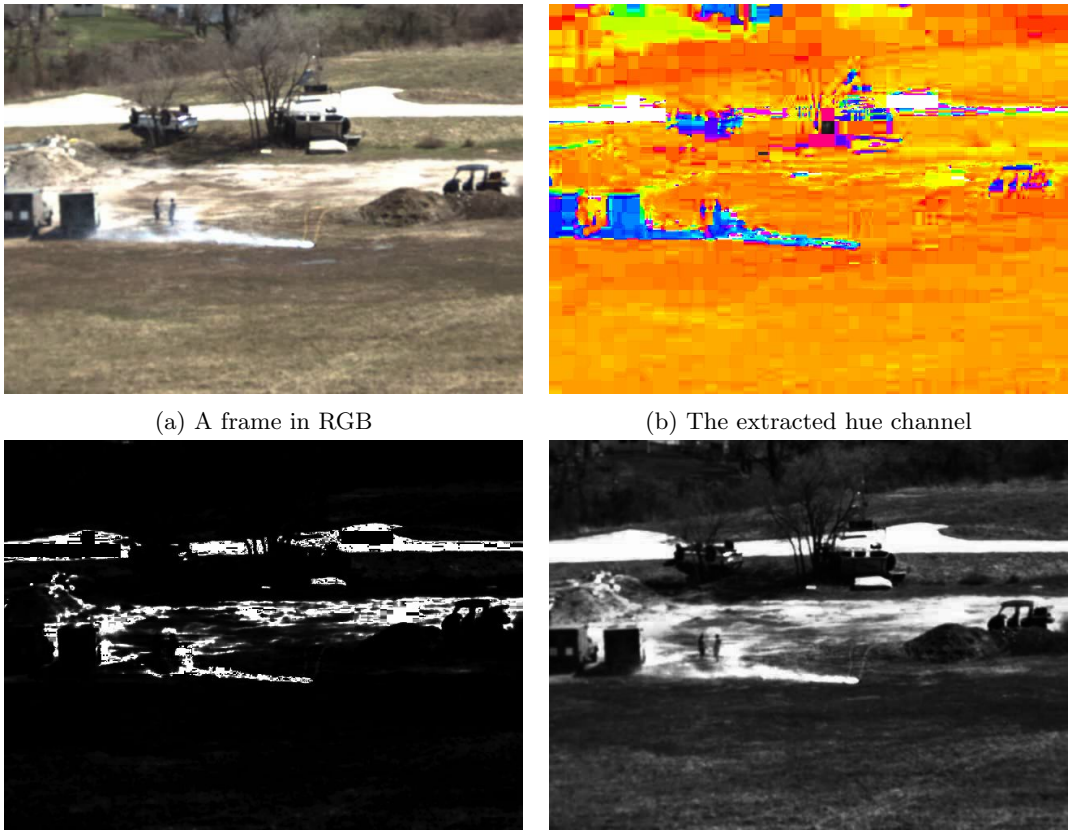
$$c = s(1 - |2l - 1|) \tag{86}$$

and

$$x = c\left(1 - \left|\frac{H}{60} \mod 2 - 1\right|\right) \tag{87}$$

the triplet (r, g, b) can be computed

$$(r,g,b) = \begin{cases} (c,x,0) & h < 60 \\ (x,c,0) & 60 \leq h < 120 \\ (0,c,x) & 120 \leq h < 180 \\ (0,x,c) & 180 \leq h < 240 \\ (x,0,c) & 240 \leq h < 300 \\ (c,0,x) & 300 \leq h \end{cases} \tag{88}$$



(a) A frame in RGB

(b) The extracted hue channel

(c) The saturation channel

(d) The lightness channel

Figure 4.11: HSL decomposition of a color frame

4.3.2 Visual Results

The decision function requires parameters k_i for each input video. By experimentation, the values were established for each modality maximizing the quality measurement as given in the next section, with $k_0 = 5.0$ (for the EO stream) and $k_1 = 6.0$ (for the IR stream).

The output video of the proposed method (depicted in Figure 4.12 is characterized by visual similarity to the input video, except in areas of occlusion by smoke. Some artifacts present in the fusion due to the scaling performed to bring the IR up to the resolution of the EO frames, as well as areas of extremely high intensity in the IR due to reflectance and weather conditions. A comparison with other off-the-shelf methods such as averaging, global DWT, and global SWT coefficient fusion is favorable for the proposed method. The target, as shown in the figure, is clearly visible through smoke.

A major difficulty in many videos, even those containing smoke, is that the target may be clearly visible through smoke in both input and result (as in Figure 4.12) due to weather conditions and smoke characteristics on site at Calamityville, thus making the optimum result very similar to the

(a) A frame from the EO input video

(b) A close-up view of the input

(c) A frame from the result video

(d) A close-up view of the region of interest

Figure 4.12: Fusion results from the proposed method

EO input. A second video was synthesized from the collected dataset with additional, artifical smoke effects. Given a mask image $B(\vec{x}) \in [0, 1]$, the EO video stream was modified according to the formula

$$I'(\vec{x}, t) = \max(I(\vec{x}, t) + B(\vec{x})U(a, b), 255) \tag{89}$$

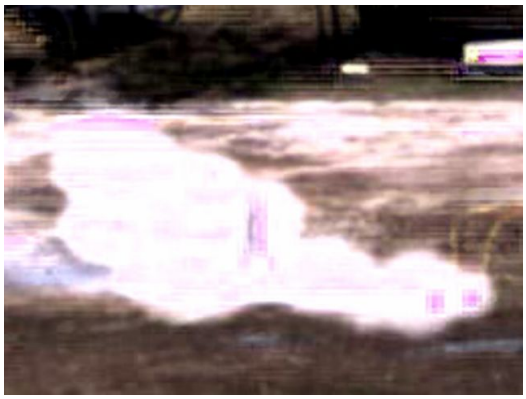
with $a = 64$, $b = 224$, and U yielding a uniform random number on the interval $[a, b]$. This allowed a region of artificial smoke-like distortion to be drawn into the image with an arbitrary shape, which can be blurred to give a more natural appearance. Output of the fusion algorithm on this second image is shown in Figure 4.13.

Visual comparison of the proposed method to averaging, global DWT fusion, and global SWT fusion is shown in Figures 4.14 and 4.15. Clipping artifacts in DWT are characteristic due to non-stationarity and ill-suited decision rules (generally mitigated through a more complex system of decision functions or resorting to more complicated wavelet bases). In SWT, low-frequency information ignored by the proposed method causes large-scale, albeit soft-edged, artifacts. A close up of concealed motion revealed by the fusion methods is shown in Figure 4.16. The proposed

(a) The mask image used

(b) Artificial smoke inserted into EO

(c) A frame from the result video



(d) A close-up view of the region of interest

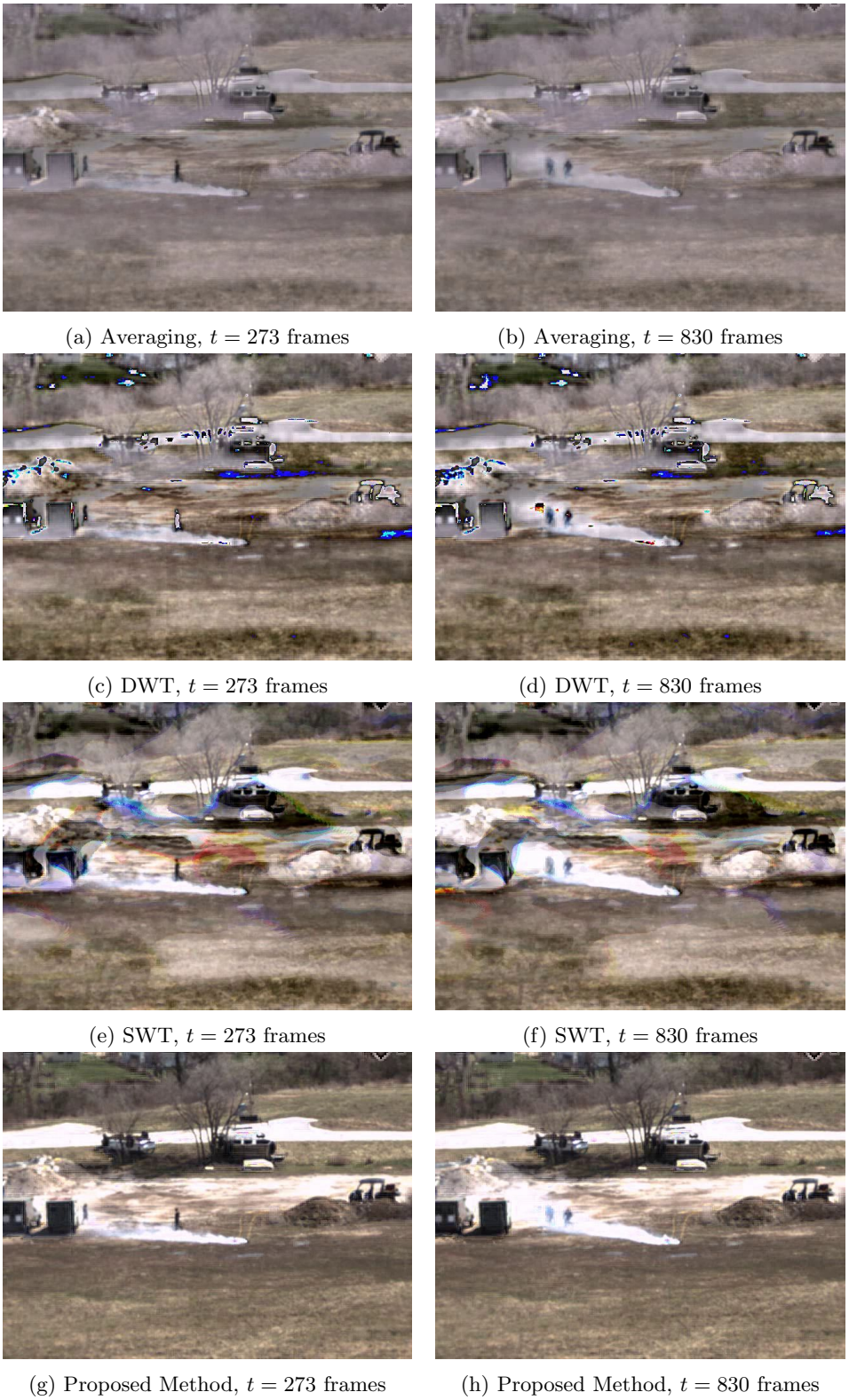
Figure 4.13: Fusion results with additional distortion

method clearly satisfies the goals given with respect to visual characteristics.

4.4 Quality Measurement

Applying Piella's quality measurement as detailed in Section 3.5 requires two implementation details: a set W of windows to analyze and a specific saliency function $s(I)$. The size of each window $w \in W$ is related to $s(I)$; w must be small enough to represent s locally, but large enough to capture an accurate result.

To analyze the videos produced by the various fusion methods mentioned in the previous section, Shannon entropy was used as a saliency function. To compute entropy, a set of symbols must be determined. If the cardinality of that set is too high, very low-amplitude noise may make a solid texture seem higher-entropy than it appears, and if the cardinality is too low, visually-distinct values may be counted as identical during comparison. The method chosen involved three steps: first, the color video was converted to grayscale. Then, the intensities within a window w (represented as bytes) were shifted down by two bits, binning them into 64 symbols. Finally, entropy was computed



(a) Averaging, $t = 273$ frames

(b) Averaging, $t = 830$ frames

(c) DWT, $t = 273$ frames

(d) DWT, $t = 830$ frames

(e) SWT, $t = 273$ frames

(f) SWT, $t = 830$ frames

(g) Proposed Method, $t = 273$ frames

(h) Proposed Method, $t = 830$ frames

Figure 4.14: Fusion results compared to other methods
The proposed method minimizes background distortion, as it avoids fusing very low-frequency data and prioritizes motion information.

(a) Averaging, $t = 273$ frames

(b) Averaging, $t = 830$ frames

(c) DWT, $t = 273$ frames

(d) DWT, $t = 830$ frames

(e) SWT, $t = 273$ frames

(f) SWT, $t = 830$ frames

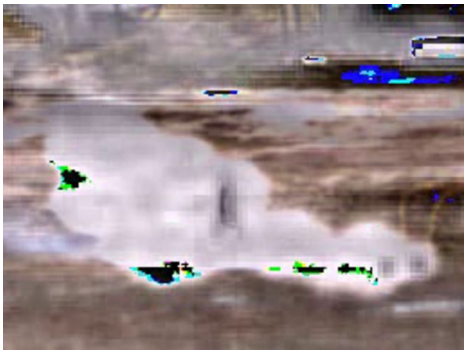
(g) Proposed Method, $t = 273$ frames

(h) Proposed Method, $t = 830$ frames

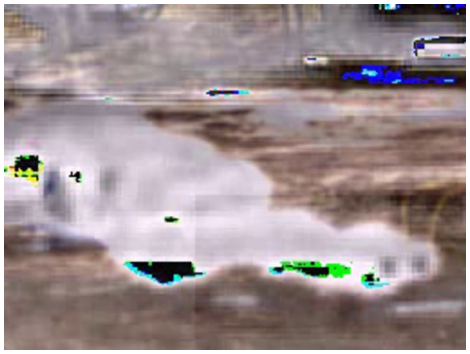
Figure 4.15: Fusion results with artificial smoke effect
Even under increased distortion, the fusion quality does not degrade, and the silhouettes are revealed without significant loss of visual quality.

(a) Averaging, $t = 273$ frames

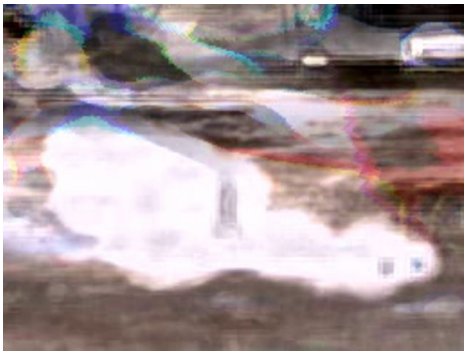
(b) Averaging, $t = 830$ frames



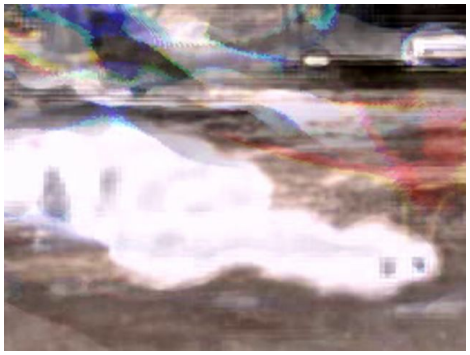
(c) DWT, $t = 273$ frames



(d) DWT, $t = 830$ frames

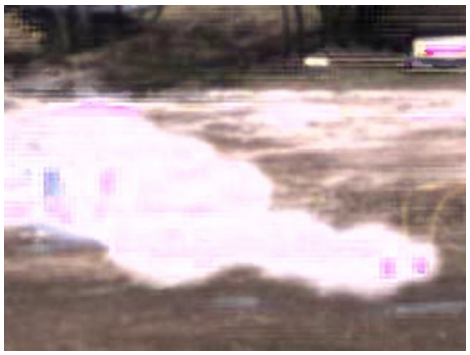


(e) SWT, $t = 273$ frames



(f) SWT, $t = 830$ frames

(g) Proposed Method, $t = 273$ frames



(h) Proposed Method, $t = 830$ frames

Figure 4.16: A closer view of fusion results across methods
The proposed method shows silhouettes under smoke, comparable to other fusion methods, but with somewhat sharper edges from incorporated motion information.

on the resulting symbols, yielding the entropy in bits as a value for $s(I|w)$. Windows of 12×12 pixels were deemed appropriate for the given videos, with W computed as a centered tiling of all 12×12 windows in the image.

Piella's formula yields a number for each fused frame, but as a basis for comparison, it suffices to take every k^{th} frame, with k proportional to the sampling rate of the video. Taking $k = 15$, we compute the mean and variance of the quality measure in the various fused videos to compare the results of the proposed method. Figure 4.17 shows the results for fusion of the original video streams and the video streams with artificial smoke inserted.

Method	μ	σ
Averaging	0.428241	0.006573
DWT	0.423233	0.006891
SWT	0.425326	0.005593
Proposed	0.453563	0.006123

(a) Q across entire image domain

Method	μ	σ
Averaging	0.359256	0.006640
DWT	0.344165	0.007455
SWT	0.352787	0.006253
Proposed	0.417339	0.007211

(b) Q in region of interest

Method	μ	σ
Averaged	0.431341	0.006617
DWT	0.422021	0.007228
SWT	0.420870	0.005517
Proposed	0.447590	0.006036

(c) Q with artificial smoke

Method	μ	σ
Averaged	0.366243	0.006970
DWT	0.341544	0.008743
SWT	0.345274	0.006086
Proposed	0.405020	0.005813

(d) Q with artificial smoke in region of interest

Figure 4.17: $Q(I_0, I_1, F(I_0, I_1))$

Piella's formula yields a number in the interval $[0, 1]$ for fusion quality. Ostensibly, this results in a zero value for a complete mismatch and a unity value for a perfectly fused image, but conceiving of those boundary cases is difficult. All of the fusion methods examined resulted in $Q < 0.5$. Further, if we examine how the unmodified input streams perform as fusion results, we discover the following:

Method	μ	σ
EO	0.446696	0.004382
IR	0.415868	0.003626
Proposed	0.453563	0.006123

(a) Q across entire image domain

Method	μ	σ
EO	0.414221	0.006887
IR	0.385594	0.004546
Proposed	0.417339	0.007211

(b) Q in region of interest

Method	μ	σ
EO	0.438631	0.004060
IR	0.426297	0.003307
Proposed	0.447590	0.006036

(c) Q with artificial smoke

Method	μ	σ
EO	0.396339	0.005432
IR	0.411733	0.004413
Proposed	0.405020	0.005813

(d) Q with artificial smoke in region of interest

Figure 4.18: $Q(I_0, I_1, I_0)$ and $Q(I_0, I_1, I_1)$ compared to fusion output

Shockingly, Piella's formula characterizes the input videos with quality values very similar to the fusion methods, and in some regions, outperforms all of them. The definition of Q_0 gives some

explanation for this:

$$Q_0(a,b) = \frac{2\mu_a\mu_b + c_1}{\mu_a^2 + \mu_b^2 + c_1} \frac{2\sigma_{ab} + c_2}{\sigma_a^2 + \sigma_b^2 + c_2} \tag{90}$$

The first term, representing mean intensity, will be constant in videos exhibiting small changes in global illumination. The second term, representing correlation and contrast similarity, is more variable among frames of the fusion techniques, but when computing $Q_0(I_0, I_0)$ or $Q_0(I_1, I_1)$, the correlation dominates, as the term reduces to a unity value. If one modality dominates in the majority of frames and regions (as is the case in the examined dataset), $Q(I_0, I_1, I_0)$ will converge towards a sum of $Q_0(I_0, I_0)$ terms, all of which are nearly constant. Finally, the fusion methods will almost always suffer when mean intensity is computed: the mean of a multimodal fusion should not necessarily resemble the frame from which information is drawn in a particular region, yet leaving the correlation term does not elevate the fusion methods above their inputs.

The alternate quality measure, Q_A, formulated in Section 3.5 gives an insight into how well a fusion method adheres to the characteristics and information of a primary modality while absorbing the additional information present in a secondary modality. Optimization over the usual computation method for $H(X, Y, Z)$ is necessary, even at small scales and with compressed intensity values. Storing an array representation of the probability density functions (PDFs) for the images is feasible with 64 symbols and 16×16 pixel windows, with joint PDFs requiring a two-dimensional matrix of 65536 values, but as a third variable is introduced, the approximately 16 million-value table begins to pose a problem for time and space complexity, especially when the number of symbols is allowed to increase. A sparse array representation is preferable for a three-variable joint PDF, requiring $O(wh)$ non-zero entries compared to the $O(|X||Y||Z|)$ entries in a full joint PDF. In this representation, a set of quadruples is stored in an array, with each three-variable observation (x, y, z) either creating a tuple $(x, y, z, 1)$, or replacing an existing tuple (x, y, z, k) with $(x, y, z, k + 1)$. The table can be further optimized by storing a list-of-lists of tuples, hashing the values x, y, z together to narrow the search space during tuple lookup. An efficient means of computing $H(X, Y, Z)$ can be easily permuted into any other joint or single-variable entropy value by summing along a row or column of the simulated three-dimensional matrix, and thus, the value Q_A is within reach, even under practical performance constraints. Results of computing Q_A on the various fusion methods, as well as pass-through fusion methods that simply copy one of the inputs into the output stream, are shown in Figure 4.19.

These results validate the visual results: the proposed method produces fused results that capture both the visual characteristics of the EO modality and the information content of the IR.

Method	μ	σ
EO	0.503338	0.000200
IR	0.431282	0.002280
Averaged	0.457488	0.008453
DWT	0.521766	0.011383
SWT	0.507575	0.006062
Proposed	0.611846	0.002808

(a) Q_A across entire image domain

Method	μ	σ
EO	0.503301	0.000292
IR	0.399908	0.003076
Averaged	0.446309	0.008702
DWT	0.494234	0.011901
SWT	0.492160	0.005962
Proposed	0.612440	0.003745

(b) Q_A in region of interest

Method	μ	σ
EO	0.503335	0.000199
IR	0.432464	0.002130
Averaged	0.466387	0.008655
DWT	0.531088	0.012419
SWT	0.519662	0.006380
Proposed	0.625767	0.003413

(c) Q_A with artificial smoke

Method	μ	σ
EO	0.503283	0.000293
IR	0.404036	0.002979
Averaged	0.467600	0.009340
DWT	0.516203	0.013156
SWT	0.519753	0.006614
Proposed	0.645117	0.005188

(d) Q_A with artificial smoke in region of interest

Figure 4.19: $Q_A(I_0, I_1, F(I_0, I_1))$

5 Significance and Contributions

The proposed methods represent three major areas of contribution: an adaptive subdivision technique for registration, spatiotemporal analysis, and a localized wavelet-domain fusion technique.

The registration technique, operating on the principle that a combination of simple registrations can yield a suitable transformation function for analysis' sake, is considerably novel, and an early form of the technique was presented at an IEEE CVPR workshop in Summer 2014 [41]. While other methods exist for performing displacement mapping or piecewise registration, the use of a subdivision method with fundamental affine transformations implies a simpler algorithm to produce comparable results to the state of the art for aerial video processing, and landmarks can be quickly discovered and evaluated via optimized filtering, feature vector computation, and extremum search techniques. The efficacy of the method is asserted in the specified use case; the method is uniquely suited to nadir-angle video containing some occlusion and varying terrain, sensitive to the level of uniformity of structure caused by different scenery.

The analysis methods are likewise novel; while segmentation in wavelet domains is well-established [42] [43], the motion analysis presented exploits an unexplored region of the problem domain. The Integral Weighted Motion technique is a novel optimization allowing an otherwise costly formula yielding a measure of unusual motion to become feasible for online use, and the wavelet-domain three frame differencing makes use of a known technique in a new way. Both represent new possibilities for efficient representation of motion in videos, and bridge the gap between temporal information and multi-scale image fusion.

Finally, the wavelet-domain fusion technique is a novel extension of an established method for image fusion. The technique reduces the amount of information lost to unnecessary fusion, producing a visually appealing result, and decreasing the amount of artifacting and false information in the output image. The proposed quality measures Q_1 and Q_A represent a new perspective on fusion, truer to the anthropocentric goals involved in multimodal fusion: the use of an objective measure that captures the visual distinctiveness of color EO imagery and the information theoretic view of supporting modalities. The methods produced can be reused in later projects to test future methods for completeness without compromising the characteristics of the host video.

6 Summary and Conclusions

Multimodal aerial imagery is subject to change in geometry, complexity, and environmental conditions. To simplify the task of an analyst in observation and decision making, an automated fusion method is proposed. The method consists of three major subsystems: an adaptive subdivision registration method, robust to changes in terrain complexity and optimized for online use, an automated analysis procedure, which characterizes structural and motion properties of each input video stream, generating comparison values, and a wavelet-domain local fusion technique, producing an optimal output video stream from multimodal video components.

Scene complexity-sensitive adaptive registration employs a number of affine transformations combined via a Gaussian basis function. By applying a course-to-fine registration, the global transformation can be repeatedly refined as the resolution of the examined regions of interest increase. The critical insight that close-up views of a projected ground plane resemble affine transformations is exploited by a system that uses increasingly detailed information on a local level, combining the resultant fundamental transformations into a smoothly interpolated global function capable of high-accuracy registration. First, landmarks are detected, then correspondence is established via optimized methods, then the obtained transformations are evaluated for further subdivision, allowing regions of high complexity to be subdivided further while not enforcing over-fitting on sparse regions. The registration that results allows for temporal analysis to take place in multimodal image streams.

Spatial and temporal analysis of multimodal videos can be performed in a multi-scale fashion, complementing the registration and fusion methods. Unlike the image fusion methods popularly used to fuse or augment multimodal videos for use by analysts, the proposed method uses motion features computed in a highly efficient manner to measure the relative importance of information across the video domain. By decomposing the input image streams via stationary wavelet transform, multiscale intensity and contrast information can be compared. Using the integral weighted motion technique, unusual motion can be quantified in the video, adding a notion of saliency to temporal analysis. Finally, for low-quality datasets, the wavelet-domain extension to the three-frame differencing technique is described, obtaining silhouettes robust to high frequency noise and changes in scene lighting. Selecting a spatial and temporal measurement module, a fusion technique can be guided to merge the information across image modalities.

Preprocessing, alignment, and calibration methods were proposed to correct for experimental factors such as imperfect camera mounting, differences in capture timing, and display range issues.

The methods given are reproducible for future data collections, with the alignment techniques given as an interactive program, and the infrared preprocessing methods easily optimized for onboard use.

With information from properly aligned image streams available, final decision and local fusion can take place. The proposed method, a stationary wavelet coefficient fusion, makes use of the already-computed decomposition of the image streams, weighting intensity and motion information across modalities, and selecting the most useful information from each input to represent in the recomposed output. The method generates a highly-intuitive video resembling the color EO input, which are validated by the quality measurement techniques suggested. By structural similarity, the output videos are highly competitive with other techniques. The additional quality measure introduced demonstrates the value of the suggested techniques, while capturing the distinctiveness of the different videos used in asymmetric, multi-modal fusion by weighing structural similarity in the primary modality against information theoretic values in the secondary modalities. Visually, objectively, and quantifiably, the output of the proposed method is validated, and the proposed fusion method is recommended for further extension and implementation in the future.

6.1 Future Work

The proposed method lends itself well to other avenues of research. A future registration method based on the proposed may make use of the Integral Weighted Motion technique to obtain regions of high subpixel inaccuracy in the background, and optimize for even greater accuracy. Seeking new techniques to guide subdivision registration may also be useful; a technique employing the arrangement of feature points, rather than their number only, may be a possible alternative.

Adaptive, automatic tuning of local contrast in the fused resuls of the proposed method may also be a feasible extension of the detailed research, with the goal of maximizing a quality measure similar to the proposed. Further, on-board optimization of the techniques may be useful, as the upcoming OpenVX architecture makes low-power deployment of an implementation of the proposed techniques possible.

References

[1] N. D. Rasmussen, B. S. Morse, M. A. Goodrich, and D. Eggett, "Fused visible and infrared video for use in wilderness search and rescue.," in *WACV*, pp. 1–8, IEEE Computer Society, 2009.

[2] Z. Zhang and R. S. Blum, "Region-Based Image Fusion Scheme For Concealed Weapon Detection," in *Proceedings of the 31st Annual Conference on Information Sciences and Systems*, pp. 168–173, 2002.

[3] A. Goshtasby, *2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications*. Wiley, 2005.

[4] E. De Castro and C. Morandi, "Registration of translated and rotated images using finite fourier transforms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, pp. 700–703, May 1987.

[5] L. Shapiro and G. Stockman, *Computer vision*. Prentice Hall, 2001.

[6] D. Gerónimo, A. D. Sappa, D. Ponsa, and A. M. López, "2d-3d-based on-board pedestrian detection system," *Comput. Vis. Image Underst.*, vol. 114, pp. 583–595, May 2010.

[7] B. S. Reddy and B. N. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration," *IEEE Transactions on Image Processing*, vol. 5, pp. 1266–1271, Aug. 1996.

[8] H. Lester and S. R. Arridge, "A survey of hierarchical non-linear medical image registration," *Pattern Recognition*, vol. 32, no. 1, pp. 129 – 149, 1999.

[9] S. Krotosky and M. Trivedi, "Multimodal stereo image registration for pedestrian detection," in *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, pp. 109 –114, 2006.

[10] R. Szeliski, "A multi-view approach to motion and stereo," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1, pp. 2 vol. (xxiii+637+663), 1999.

[11] C. Dai, Y. Zheng, and X. Li, "Layered representation for pedestrian detection and tracking in infrared imagery," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops - Volume 03*, CVPR '05, (Washington, DC, USA), pp. 13–, IEEE Computer Society, 2005.

[12] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2, pp. 2 vol. (xxiii+637+663), 1999.

[13] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2, pp. 28 – 31 Vol.2, aug. 2004.

[14] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, "Background modeling and subtraction of dynamic scenes," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 1305 –1312 vol.2, oct. 2003.

[15] E. Hayman and J.-O. Eklundh, "Statistical background subtraction for a mobile observer," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 67 –74 vol.1, oct. 2003.

[16] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: principles and practice of background maintenance," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1, pp. 255 –261 vol.1, 1999.

[17] O. Javed, K. Shafique, and M. Shah, "A hierarchical approach to robust background subtraction using color and gradient information," in *Motion and Video Computing, 2002. Proceedings. Workshop on*, pp. 22–27, Dec 2002.

[18] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Comput. Vis. Image Underst.*, vol. 106, pp. 288–299, May 2007.

[19] Y. Fang, K. Yamada, Y. Ninomiya, B. Horn, and I. Masaki, "A shape-independent method for pedestrian detection with far-infrared images," *IEEE Transactions On Vehicular Technology*, vol. 53, no. 5, p. 1679, 2004.

[20] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM Journal of Mathematical Analysis*, vol. 15, no. 4, pp. 723–736, 1984.

[21] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, 1989.

[22] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform," 1989.

[23] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," pp. 281–300, Springer-Verlag, 1995.

[24] J. Magarey and N. Kingsbury, "Motion estimation using a complex-valued wavelet transform," *Signal Processing, IEEE Transactions on*, vol. 46, no. 4, pp. 1069–1084, 1998.

[25] N. Kingsbury, "The dual-tree complex wavelet transform: A new technique for shift invariance and directional filters," pp. 319–322.

[26] J. J. Lewis, R. J. O. Callaghan, S. G. Nikolov, and D. R. Bull, "Region-based image fusion using complex wavelets," *Image Rochester NY*, vol. 8, no. 2, pp. 119–130, 2007.

[27] A. P. Bradley, "Shift invariance in the discrete wavelet transform," in *In VIIth Digit. Image Comp*, vol. 4, pp. 29–38, 2003.

[28] M. I. Restrepo, B. A. Mayer, A. O. Ulusoy, and J. L. Mundy, "Characterization of 3-d volumetric probabilistic scenes for object recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, pp. 522–537, 2012.

[29] M. I. Restrepo, *Characterization of Probabilistic Volumetric Models for 3-D Computer Vision*. PhD thesis, Brown University, 2013.

[30] M.-K. Hu, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions on*, vol. 8, pp. 179–187, February 1962.

[31] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, June 1981.

[32] A. Goshtasby, "A weighted linear method for approximation of irregularly spaced data.." Lucian, Miriam L. (ed.) et al., Geometric modeling and computing: Seattle 2003. Selected papers of the 8th SIAM conference on geometric design and computing, Seattle, WA, USA, November 9–13, 2003. Brentwood, TN: Nashboro Press. Modern Methods in Mathematics, 285-294 (2004)., 2004.

[33] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Hawaii), 2001.

[34] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, pp. 81–84, Mar. 2002.

[35] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *Geoscience and Remote Sensing Letters, IEEE*, vol. 1, no. 4, pp. 313–317, 2004.

[36] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 3, pp. III–173, IEEE, 2003.

[37] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. S. Crnojevic, and D. Kukolj, "Salient motion features for video quality assessment.," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 948–958, 2011.

[38] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[39] D. Blostein and N. Ahuja, "A multiscale region detector," *Computer Vision, Graphics, and Image Processing*, vol. 45, no. 1, pp. 22–41, 1989.

[40] W. J. Carper, T. M. Lillesand, and P. W. Kiefer, "The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data," *Photogrammetric Engineering and Remote Sensing*, vol. 56, pp. 459–467, 1990.

[41] B. Jackson and A. Goshtasby, "Adaptive registration of very large images," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pp. 351–356, June 2014.

[42] G. Piella, "A general framework for multiresolution image fusion: from pixels to regions," *Information Fusion*, vol. 4, pp. 259–280, Dec. 2003.

[43] J. Li and Y. Wang, "Pedestrian tracking in infrared image sequences using wavelet entropy features," in *Computational Intelligence and Industrial Applications, 2009. PACIIA 2009. Asia-Pacific Conference on*, vol. 1, pp. 288–291, 2009.