A Multi-Sensor Fusion Framework in 3-d

Vishal Jain
Vision Systems Inc.
Providence, RI
vishal@visionsystemsinc.com

Andrew C. Miller
Dept. of Computer Science
Harvard University
acm@seas.harvard.edu

Joseph L. Mundy
Vision Systems Inc.
Providence, RI
mundy@lems.brown.edu

Abstract

The majority of existing image fusion techniques operate in the 2-d image domain which perform well for imagery of planar regions but fails in presence of any 3-d relief and provides inaccurate alignment of imagery from different sensors. A framework for multi-sensor image fusion in 3-d is proposed in this paper. The imagery from different sensors, specifically EO and IR, are fused in a common 3-d reference coordinate frame. A dense probabilistic and volumetric 3-d model is reconstructed from each of the sensors. The imagery is registered by aligning the 3-d models as the underlying 3-d structure in the images is the true invariant information. The image intensities are back-projected onto a 3-d model and every discretized location (voxel) of the 3-d model stores an array of intensities from different modalities. This 3-d model is forward-projected to produce a fused image of EO and IR from any viewpoint.

1. Introduction

Multi-sensor fusion techniques combine data from multiple sensors to achieve improved accuracies and more specific inferences than could be achieved by the use of a single sensor alone. Such techniques are widely used in both military and commercial applications, such as target-detection, tracking and land classification for remote sensing. Specifically, multi-sensor fusion facilitates (i) improved confidence in decisions, such as target location and identity, by fusing different sources of information. For example geo-coordinates from GPS and high resolution Electro-optical (EO) imagery , can be fused with a distinctive signature from Infra-Red (IR) imagery for identification; (ii) improved target detection especially in the case of countermeasures (camouflage, cluttered scenes, etc.) which can be defeated by exploiting a range of spectral bands, or even polarization sensing; (iii) robust performance by providing extended range of operating conditions and improved performance under adverse environmental conditions, *e.g.*, low visibility due to smoke or fog for EO can be mitigated by IR or GMTI.

The majority of existing image fusion techniques operate in the 2-d image domain [1]. Typically, images from different sensors are registered by estimating a mapping that aligns scene features between 2-d images, but these techniques fail in the presence of 3-d relief, roof tops of the buildings are misaligned as shown in Figure 1 c). Aligning two images may seem straightforward since humans can perceive 3-d structures in images accurately due to their contextual information and recognition abilities. However, computer systems perceive these images as 2-d arrays of intensities and are unable to directly relate 3-d structures in one image to another. Approaches such as [2], [3] register 2-d images to existing 3-d models for accurate alignment of 3-d relief in the images. However, the estimation of an alignment of a 2-d image to a 3-d model suffers from projective ambiguity, i.e., multiple transformations can align the image with the 3-d model. Additionally, there is the lack of availability of high-resolution and up-to-date 3-d models.

This paper presents a framework for multi-sensor image fusion in 3-d. The imagery from different sensors, specifically EO and IR, are fused in a common 3-d reference coordinate frame. The images from a single sensor are used to build a dense probabilistic and volumetric 3-d model [4–6] corresponding to each sensor. These dense volumetric models are of the same resolution as the imagery. The individual 3-d models from each sensor are registered to each other via existing 3-d to 3-d matching [7][8][9]. By registering the 3-d models, the corresponding imagery is also registered automatically. The image intensities are back-projected onto a 3-d model corresponding to the highest resolution as shown in Figure 1 d) and e). Every discretized location of the 3-d model stores an array of intensities from different modalities. This 3-d model can be forward-projected to produce a fused image from any viewpoint, Figure 1 f).

A major application of multi-sensor fusion is extended surface attribution for material and object classification. Information from different sensors, when fused, provides a signature for object surfaces. Multispectral data constructed from fused video streams can be processed with a wide range of material classification algorithms developed by the remote sensing community over the last

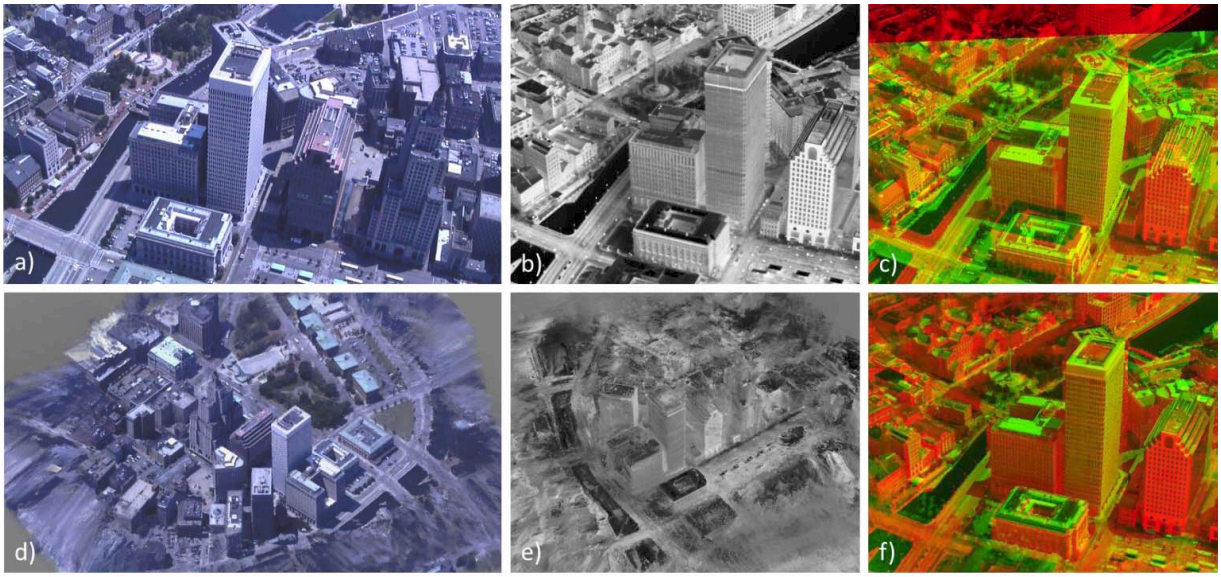

Figure 1. Fusion of EO and IR in 3-d as compared to 2-d. a) and b) show EO and IR images, respectively, to be fused; c) shows the results of a 2-d registration approach using a projective homography. The red channel is used for the IR image and the green is used for intensity of EO image. Note that the registration is not accurate at the top of the buildings. d) and e) 3-d models constructed from color EO and IR imagery, respectively. f) shows the rendering of the 3-d fused model from a viewpoint similar to c). Note that the imagery is accurately aligned.

few years [10]. The effectiveness of the 3-d fusion framework is further illustrated by evaluating region classification in images generated by the fused 3-d model. A machine learning technique, Support Vector Machines (SVM), is used to train the classifier on the fused imagery and to classify the fused imagery of the test data into different regions.

2. Related Work

Multi-sensor fusion is prevalent in numerous fields such as ocean engineering, navigation, robotics and controls, computer vision, artificial intelligence and many more. In all of these fields, measurements from numerous sensors need to be fused to derive robust estimates and inferences. Hall, [1], lays out a complete hierarchical fusion framework for various layers of information ranging from raw sensor measurements to high level inferences. This paper describes different fusion techniques used for various sources of information in the presence of noise and bias. This work demonstrates that fusion is essential not only at the raw sensor data level but also for high–level information, such as target detection inferences from the measurements. DARPA's urban challenge, [11], is a good illustration of these concepts where measurements from a large amount of external sensors are fused to extract high-level situation awareness information for control and decision making in autonomous driving.

There exists a multitude of 2-d based fusion approaches where the images from EO and IR sensors are registered in the image domain. Since it is not practical to review all the existing approaches here, of the most relevant approaches are described. These approaches are applicable to planar scenes, with sufficiently large viewing distance so that a 2-d affine transformation is sufficient to align images of different modalities. The authors in [12] propose to align the images from multiple EO and IR sensors. An image-to-image affine transformation is estimated by aligning the gradient maps. The hypothesis is the gradient of the images is more invariant across different sensors. Another variant of the above approach, [13], uses the gradient along with the gradient vector field to find mapping between visible and infrared images. Another set of approaches, [14], [15], uses foreground image segmentations and their motion trajectories to register imagery from different modalities. Numerous feature-based approaches exist such as corner-detection based methods [16], [17] and wavelet-description based methods [18] for image alignment.

The most relevant and closely related 2-d to 3-d registration approach is proposed in [2]. The authors use 3-d site models to register imagery from different airborne sensors. The images are registered to a 3-d site model and the intensities from the imagery are texture-mapped onto the site model via a color fusion process. The site-models used in the work are, for the most part, widely available coarse resolution Depth Elevation Models (DEMs) or manually extracted 3-d models. The limitations of this approach are (i) that the 2-d to 3-d registration typically suffers from projective ambiguity which leads to inaccurate transformation between the 2-d image and 3-d model, (ii) lack of high-fidelity and up-to-date models for registering imagery.

Another relevant approach, [3], uses dense octree-based volumetric models created from engineering designs or the

silhouettes of objects from the imagery to predict multi-sensor imagery. This approach is useful to study the physical models of the multi-spectral sensors, but is not robust to the complexity of real-world aerial video streams By contrast, the approach presented in this paper can infer 3-d models directly from airborne imagery from each sensor and register the imagery reliably using 3-d to 3-d registration which is free of any projection ambiguity. The resulting multi-spectral model is capable of supporting many analytic services such as predicting multi-spectral surface properties.

3. Multi-Sensor Fusion in 3-d

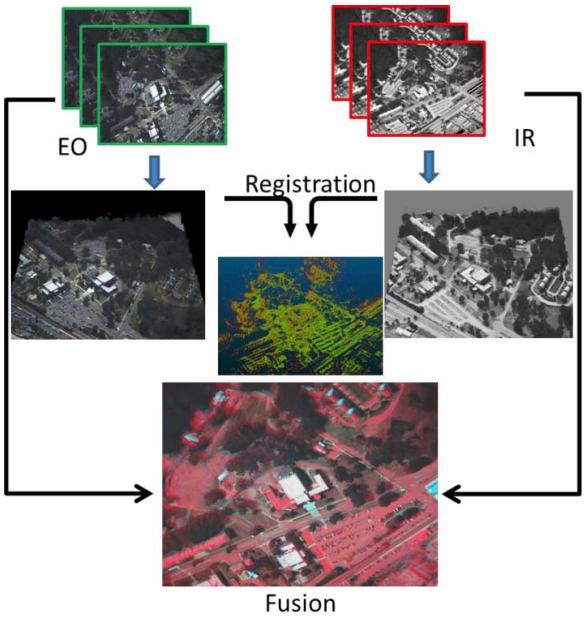

Figure 2 A schematic diagram of the proposed approach for fusion of EO and IR imagery in 3-d.

The proposed approach, Figure 2, is novel in that it registers images in 3-d, which allows for an accurate and robust alignment of multi-sensor imagery. This approach reconstructs a probabilistic dense 3-d model, using [5], for each modality EO and IR separately. VisualSFM [6] is used to estimate the intrinsic and extrinsic camera parameters for each modality independently. The dense volumetric 3-d model is built for each modality from both the imagery and the associated camera parameters as shown in Figure 4. At this stage, the two models have independent coordinate systems and need to be registered to a common coordinate system. The two dense 3-d models, EO and IR, are converted into point clouds and are registered using a feature-based alignment algorithm followed by few iterations of Iterative Closest Point (ICP) algorithm [9].

The transformation obtained from registering the two 3-d models is used to transform the projection parameters (cameras) of the imagery. The 3-d model with higher resolution, typically EO, is chosen as a reference model. The EO and IR imagery with registered cameras is back-projected into the model and is stored at each discretized 3-d location or voxels, as modulated by voxel surface probability and visibility The 3 RGB channels from EO and the grayscale value from IR are stored at each voxel either as 4 channels or fused together using a color conversion function, where IR is the red channel and the green and blue channels are used from EO. This results in a 3-d model that can be rendered from any viewpoint to obtain fused imagery. Each of the above steps is discussed in detail in the following subsections.

3.1. Overview of 3-d modeling

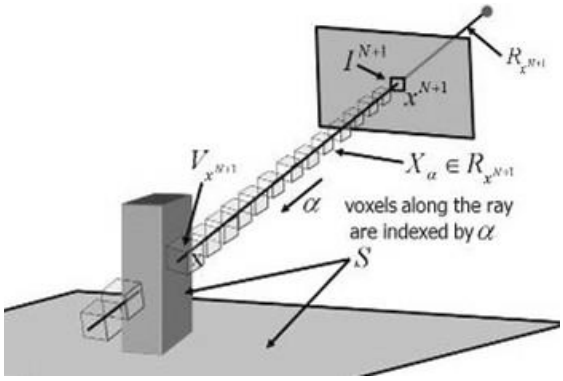

Figure 3 The interaction of a ray with the cells in the volume for probabilistic inference such as rendering image.

Probabilistic Volumetric Representation (PVR) technology, [4], [5] , is used to build 3-d models from EO and IR imagery automatically as shown in Figure 4 b) and d). The PVR is a volumetric dense 3-d grid of voxels where each voxel contains a probability distribution for surface geometry and the associated appearance of each surface element. A voxel is assigned to one of two states: a surface, S, or not a surface. The belief of a voxel, X, being a surface is denoted by a probability, $P(X \in S)$. The appearance of each voxel is represented by a probability density $P(I|X \in S)$ which is modeled with a Mixture of Gaussians (MoG). Also note that the $P(I|X \notin S)$ is undefined - the appearance of a voxel given that it is not a surface.

To reconstruct the geometry and appearance of a volume, each voxel incorporates pixel information from a set of images (video frames, satellite images, *etc.*). Each pixel of an input image, I, is back-projected into the volume along the ray, R, using a known camera, shown in Figure 3. The underlying assumption is that only one voxel V_x along ray R produces the observed intensity in the image. The voxel must be un-occluded and have a high surface probability. This condition will be referred to as the ``responsibility" of a voxel producing some intensity in image I, which is calculated as

$$P(x = V_x) = P(X \in S) \prod_{x' < x} (1 - P(x' \in S))$$

where the product range $x' < x$ indicates the voxels closer to the sensor than voxel x. Each voxel also has an appearance distribution, $P(I|X \in S)$. Given the responsibility and appearance of voxels along a ray, the expected intensity of some ray R cast through the model can be computed as

$$E[I] = \sum_{x \in R} P(x = V_x) E[P(I|X \in S)],$$

where the range $x \in R$ indicates the set of voxels that intersect with ray R, ordered from closest to farthest from the camera center.

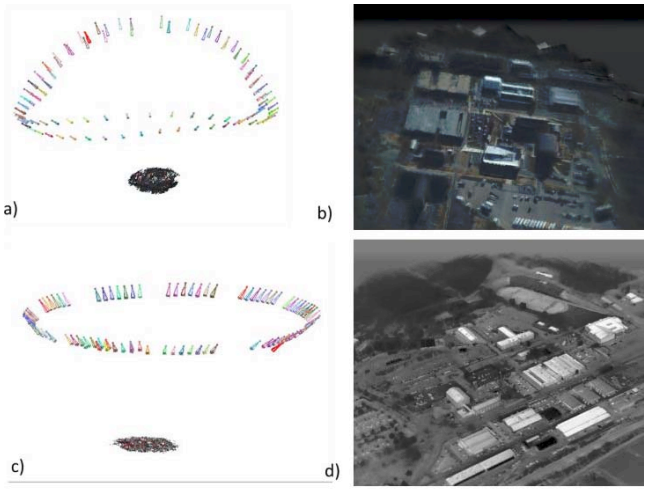

Figure 4 3-d models reconstructed from video sequence of EO and IR images. a) and c) Sensor positions and viewpoints estimated using [6] **for EO and IR images, respectively. b) and d) Renderings of a 3-d model constructed from EO and IR image, respectively.**

The online updating algorithm for each process is a consequence of Bayes' theorem. The updating of the surface probability of voxel X given the input image intensity i is given by

$$p^{N+1}(X \in S) = p^N(X \in S) \frac{p^N(I|X \in S)}{P^N(I)},$$

where p^N indicates the distribution at time N. This equation intuitively states that if the appearance model of voxel x accurately explains observed intensity I, the probability of that voxel being a surface is increased. Whereas if the appearance model of voxel x assigns a low probability density to intensity i, the probability of voxel x being a surface is decreased. The parameters of the Mixture of Gaussian (MoG) for voxel X are updated given the input image intensity I using equations from [19].

The probability distribution of the surface geometry facilitates modeling of uncertainty or ambiguities in 3-d terrain models, as opposed to mesh-based reconstruction techniques cannot handle any uncertainty. The PVR technology uses octrees for efficient storage, [20], and GPU implementation, [5], for efficient computation. The input to PVR technology is a sequence of pairs of an image and its corresponding camera to build the dense 3-d model automatically. Typically, PVR has been extensively used to build models from EO imagery but this work also demonstrates 3-d models reconstructed from IR imagery as well. A PVR 3-d model can be rendered from a novel viewpoint to generate an image and the 3-d model surfaces can be painted with different intensities.

3.2. 3-d to 3-d Registration

A 3-d to 3-d registration approach is used to register 3-d models reconstructed from each set of EO and IR image sequences. Unlike, image-to-image or image-to-3d model registration approaches, 3-d to 3-d registration intrinsically accounts for 3-d terrain relief and is not affected by intensity variations, even across different modalities. The underlying 3-d terrain is completely independent of how it is sensed.

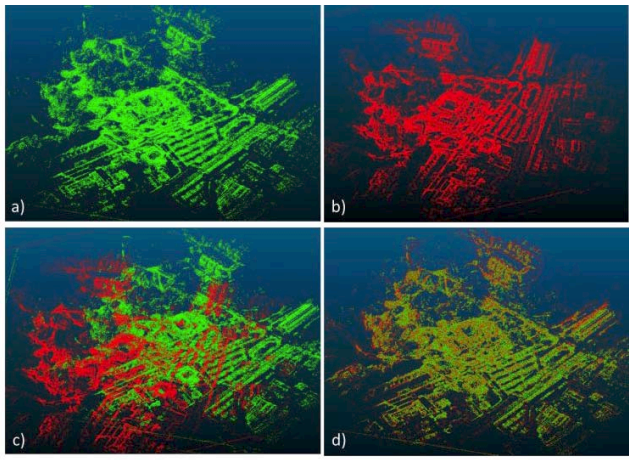

Figure 5: 3-d to 3-d registration by aligning the 3-d point clouds. a) and b) point clouds obtained from dense volumetric 3-d models for EO and IR imagery, respectively. c) initial alignment of the two point clouds and d) accurate alignment after applying [7], [9].

There exist numerous approaches for registering or aligning 3-d point clouds. The Iterative Closest Points (ICP) algorithm [9], and its variants, are the most commonly used algorithm for registering 3-d point sets. Such algorithms do provide accurate registration, however they require the point-clouds to initially be within close proximity. If this condition is not met, feature-based approaches [7] are used to align 3-d point clouds which differ by large rotations and translation. The dense 3-d probabilistic volumetric models are converted into a 3-d point cloud by applying a threshold to the surface probabilities of the voxels.

The 3-d point clouds obtained from dense 3-d models of EO and IR imagery are registered by estimating a similarity transform using [7]. The approach in [7]

computes a histogram of relative orientations of the neighboring points of the location under consideration. This feature is robust and discriminative and a Sample Consensus Initial Alignment is used to estimate the best transformation by minimizing the distance between matching features. This approach is typically able to estimate an alignment for the two point-clouds which is further refined using ICP, [9], to ensure high accuracy registration, Figure 5.

3.3. Fusion of EO and IR Imagery

The biggest advantage of fusing EO and IR imagery in 3-d is the accurate alignment of the images from different sensors in spite of the presence of occlusion due to the 3-d structures present in the scene. Thus, the imagery from different sensors does not need to be collected from the same viewpoint, since the images from very different viewpoints can be accurately fused.

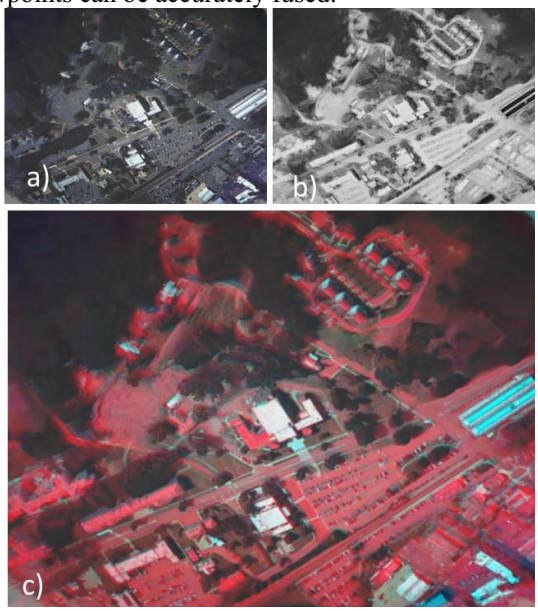

Figure 6 a) A frame from an IR video sequence; b) a frame from an EO sequence; c) shows the fused EO and IR image.

The aligned imagery from EO and IR sensor using the 3-d registration approach is fused in 3-d. The 3-d model corresponding to the higher resolution modality is chosen as the reference model for fusion, denoted by M. The color EO images have typically three channels, RGB, and IR images have a single channel. These four channels are converted into 3 channels using a transform, where IR is the red channel and the green and blue channels are used from EO, for visualization.

For fusing two image sequences that are synchronized or unsynchronized, denoted by $\{F_1^i\}^m$ and $\{F_2^i\}^n$ which have m and n images respectively. Each of the image sequences update its respective appearance models at 3-d locations of the reference model, M. The intensities from all the images from a single modality are accumulated in the same appearance model at each location, Figure 6 a) and b). This allows for either direct use of the 3-d model itself or the rendered 2-d images different viewpoints as shown in Figure 6 c).

4. Scene Classification

The effectiveness of 3-d fusion framework is further illustrated by classifying different pixels from fused EO and IR imagery. A set of fused rendered images from a 3-d model of the urban Virginia dataset, Figure 7 a), are used for training. A 14-dimesnional feature vector for each pixel in the training set is estimated. This feature vector comprises of raw intensities of all the channels, 4, pairwise difference of intensities in different channels, 6, and ratio of each channel value to the sum of all the channel intensities, 4. A Support Vector Machine (SVM) classifier is trained on manually segmented pixels from 5 rendered images of the fused model from a site in Virginia, Figure 7 a). The SVM classifier is used to classify pixels in the rendered images of the fused model from a different site in Virginia, Figure 7 b) and the result is shown in Figure 7 d). The different categories are assigned different colors: trees (green), parking lots (blue), roads (magenta) and buildings (red). Note that black color is used for no category.

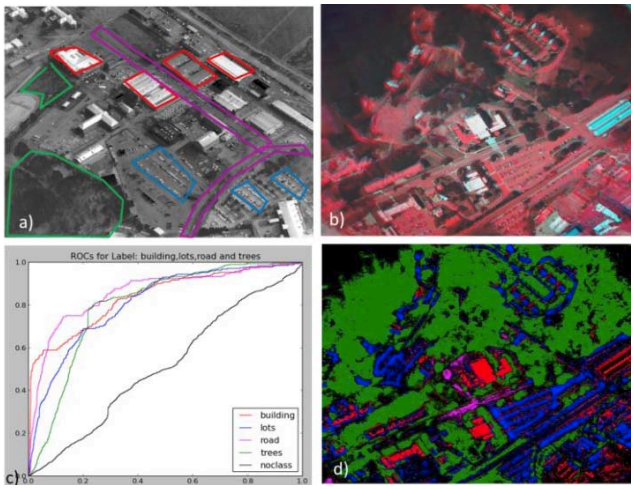

Figure 7 a) Manually labeled regions for training the classifier; b) Clusters for different learned categories trees (green), parking lots (blue), roads (magenta), buildings (red) and no-class (black). c) ROC curve for different classes; and d) the region classification.

The classification of rendered images the fused 3-d model is also evaluated quantitatively. A ROC curve is plotted for each category and the performance of all the categories was satisfactory except that of the no-class category. The no-class category is a very broad category and requires extensive training data. Due to the limited training data, the performance of the no-class category is close to random. Furthermore, the pixel classification of the 3-d fused image, Figure 8 a), was compared to the classification on a 2-d fused image, Figure 8 b). The latter image shows poor pixel classification in areas of mis-

registration such as building tops. Also, the classification using EO image and IR image alone are shown in Figure 8 c) and d) respectively. It can be seen that both EO and IR images individually are able to classify only few of the categories.

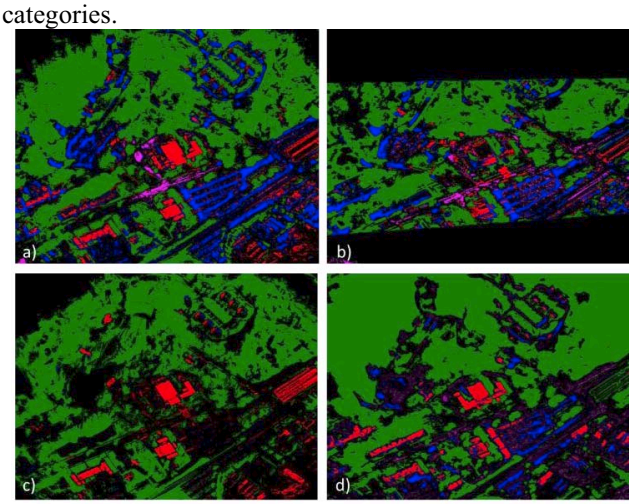

Figure 8 compares the pixel classification of the a) 3-d fused image, b) 2-d fused image, c) EO image alone and d) IR image alone.

5. Conclusion

A multi-sensor fusion framework for EO and IR imagery in 3-d is proposed. Unlike, existing 2-d approaches, which cannot accurately align imagery in presence of 3-d relief, the proposed framework accurately aligns the imagery.

References

[1] J. Llinas and D. L. Hall, "An introduction to multi-sensor data fusion," *Circuits and Systems 1998 ISCAS98 Proceedings of the 1998 IEEE International Symposium on*, vol. 6, pp. 537–540, 2002.

[2] W. D. Ross, A. M. Waxman, W. W. Streilein, M. Aguiiar, J. Verly, F. Liu, M. I. Braun, P. Harmon, and S. Rak, *Multi-sensor 3D image fusion and interactive search*, vol. 1. 2000, pp. 10–17.

[3] J. Michel and N. Nandhakumar, *Unified 3D models for multisensor image synthesis*, vol. 57, no. 4. Academic Press Inc JNL-Comp subscriptions, 1994, pp. 283–302.

[4] T. Pollard and J. L. Mundy, *Change Detection in a 3-d World*, no. C. Ieee, 2007, pp. 1–6.

[5] A. Miller, V. Jain, and J. L. Mundy, "Real-time rendering and dynamic updating of 3-d volumetric data," *Proceedings of the Fourth Workshop on General Purpose Processing on Graphics Processing Units GPGPU4*, p. 1, 2011.

[6] W. Changchang, "VisualSFM: A Visual Structure from Motion System." 2011.

[7] R. B. Rusu, N. Blodow, and M. Beetz, "Fast Point Feature Histograms (FPFH) for 3D registration," *Proceedings of the IEEE International Conference on Robotics and Automation (2009)*, pp. 3212–3217, 2009.

[8] W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information.," *Medical Image Analysis*, vol. 1, no. 1. Elsevier, pp. 35–51, 1996.

[9] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[10] Y. Her and G. Student, "Land use classification in Zambia using Quickbird and Landsat imagery," *Methods*, vol. 0300, no. 07, 2007.

[11] M. Buehler, K. Iagnemma, and S. Singh, *Journal of Field Robotics: Special Issue on the 2007 DARPA Ur ban Challenge, Part I–III*. 2008.

[12] J. K. J. Kang, K. Gajera, I. Cohen, and G. Medioni, *Detection and Tracking of Moving Objects from Overlapping EO and IR Sensors*, vol. 00, no. C. Ieee, 2004, pp. 123–123.

[13] J. H. L. J. H. Lee, Y. S. K. Y. S. Kim, D. L. D. Lee, D.-G. K. D.-G. Kang, and J. B. R. J. B. Ra, *Robust CCD and IR Image Registration Using Gradient-Based Statistical Information*, vol. 17, no. 4. 2010, pp. 347–350.

[14] G. A. Bilodeau, A. Torabi, and F. Morin, "Visible and infrared image registration using trajectories and composite foreground images," *Image and Vision Computing*, vol. 29, no. 1, pp. 41–50, 2011.

[15] Y. Caspi, D. Simakov, and M. Irani, "Feature-Based Sequence-to-Sequence Matching," *International Journal of Computer Vision*, vol. 68, no. 1, pp. 53–64, 2006.

[16] C. Park, K. Bae, S. Choi, and J.-H. Jung, "Image fusion in infrared image and visual image using normalized mutual information," in *Proceedings of SPIE Vol 6968*, 2008, vol. 6968, no. 1, p. 69681Q–69681Q–9.

[17] J.-W. Z. J.-W. Zhang, G.-Q. H. G.-Q. Han, and Y. W. Y. Wo, *Image registration based on generalized and mean Hausdorff distances*, vol. 8. 2005.

[18] H. X. H. Xishan and C. Z. C. Zhe, *A wavelet-based multisensor image registration algorithm*, vol. 1. Ieee, 2002, pp. 773–776.

[19] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *Proceedings 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Cat No PR00149*, vol. 2, no. c, pp. 246–252, 1999.

[20] D. Crispell, J. Mundy, and G. Taubin, "A Variable-Resolution Probabilistic Three-Dimensional Model for Change Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 489–500, 2012.