UNIVERSITY OF CALIFORNIA, SAN DIEGO

On Using Multiperspective Color and Thermal Infrared Videos to Detect People: Issues, Computational Framework, Algorithms and Comparative Analysis

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in
Electrical Engineering (Signal and Image Processing)

by

Stephen Justin Krotosky

Committee in charge:

Professor Mohan M. Trivedi, Chair
Professor Serge Belongie
Professor Pamela Cosman
Professor Vistasp Karbhari
Professor Truong Nguyen

2007

The dissertation of Stephen Justin Krotosky is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2007

So I can see you smile.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge my family, friends and colleagues without whose help and support this dissertation would not be possible.

I would firstly like to give thanks to my advisor, Professor Mohan M. Trivedi, whose encouragement and enthusiasm has been invaluable in inspiring me to develop and pursue my independent research. His careful advisement and suggestions have greatly helped to improve this dissertation.

I would also like to thank the U.S. Department of Defense Technical Support Working Group for Combatting Terrorism and the UC Discovery Program whose sponsorship has allowed me to freely pursue this area of research.

In addition, I would like to express my gratitude to my doctoral committee: Professor Serge Belongie, Professor Pamela Cosman, Professor Vistasp Karbhari and Professor Truong Nguyen. Thank you for your time and valuable inputs that have allowed me to reach my research and academic goals.

I owe a special debt of thanks to my labmates in the CVRR lab for their support and friendship, specifically Shinko Cheng, Anup Doshi, Dr. Tarak Gandhi, Dr. Kohsia Huang, Ramsin Khoshabeh, Dr. Joel McCall, Brendan Morris, Erik Murphy-Chutorian, Dr. Sangho Park, Shankar Shivappa, and Dr. Junwen Wu. Your continued willingness to participate in data collection and indulge my research questions and discussion has helped me through some of the more difficult stretches of my research.

I would like to thank my parents for their support and encouragement in pursuing academic endeavors.

Finally, I would like to thank Mychang Van for her unwavering encouragement and boundless love. I would not have been able to devote the days and days of time and effort in lab without the knowledge that your support and love would welcome me home at night. I dedicate this dissertation to you.

La Jolla *Stephen Krotosky*

August 28, 2007.

The text of Chapter II, in part, is a reprint of the material as it appears in: Stephen J. Krotosky and Mohan M. Trivedi, "Mutual Information Based Registration of Multimodal Stereo Videos for Person Tracking" in *Computer Vision and Image Understanding, Special Issue on Advances in Vision Algorithms and Systems Beyond the Visible Spectrum*, Vol. 106, Issues 2-3, May-June 2007. I was the primary researcher of the cited material and the co-author listed in this publication directed and supervised the research which forms a basis for this chapter.

The text of Chapter III, in part, is a reprint of the material as it appears in: Stephen J. Krotosky and Mohan M. Trivedi, "Mutual Information Based Registration of Multimodal Stereo Videos for Person Tracking" in *Computer Vision and Image Understanding, Special Issue on Advances in Vision Algorithms and Systems Beyond the Visible Spectrum*, Vol. 106, Issues 2-3, May-June 2007 and Stephen J. Krotosky and Mohan M. Trivedi, "Registration of Multimodal Imagery with Occluding Objects using Mutual Information", *Applied Perception in Thermal Infrared Imagery*, in press. I was the primary researcher of the cited material and the co-author listed in these publications directed and supervised the research which forms a basis for this chapter.

The text of Chapter IV, in part, is a reprint of the material as it appears in: Stephen J. Krotosky and Mohan M. Trivedi, "Mutual Information Based Registration of Multimodal Stereo Videos for Person Tracking" in *Computer Vision and Image Understanding, Special Issue on Advances in Vision Algorithms and Systems Beyond the Visible Spectrum*, Vol. 106, Issues 2-3, May-June 2007 and Stephen J. Krotosky and Mohan M. Trivedi, "On Color, Infrared and Multimodal Stereo Approaches to Pedestrian Detection", *IEEE Trans. On Intelligent Transportation Systems*, in press. I was the primary researcher of the cited material and the co-author listed in these publications directed and supervised the research which forms a basis for this chapter.

The text of Chapter V, in part, is a reprint of the material as it appears in: Stephen J. Krotosky and Mohan M. Trivedi, "On Color, Infrared and Multimodal Stereo Approaches to Pedestrian Detection", *IEEE Trans. On Intelligent Transportation Systems*, in press. I was the primary researcher of the cited material and the co-author listed

in this publication directed and supervised the research which forms a basis for this chapter.

The text of Chapter VI, in part, is a reprint of the material as it appears in: Stephen J. Krotosky and Mohan M. Trivedi, "Algorithmic Framework and Experimental Evaluation for using Multiperspective Color and Infrared Features for Person Surveillance", *IEEE Trans. on Circuits and Systems for Video Technology*, submitted. I was the primary researcher of the cited material and the co-author listed in this publication directed and supervised the research which forms a basis for this chapter.

VITA

2001	B.S. Computer Engineering University of Delaware
2002-2007	Graduate Student Researcher University of California, San Diego
2004	M.S. Electrical Engineering (Signal and Image Processing) University of California, San Diego
2007	Ph.D. Electrical Engineering (Signal and Image Processing) University of California, San Diego

PUBLICATIONS

On Color, Infrared and Multimodal Stereo Approaches to Pedestrian Detection
S. J. Krotosky and M. M. Trivedi, *IEEE Trans. on Intelligent Transportation Systems*, In Press.

A Comparison of Color and Infrared Stereo Approaches to Pedestrian Detection
S. J. Krotosky and M. M. Trivedi, *Proc. IEEE Intelligent Vehicles Symposium*, June 2007.

Mutual Information Based Registration of Multimodal Stereo Videos for Person Tracking
S. J. Krotosky and M. M. Trivedi, *Mutual Information Based Registration of Multimodal Stereo Videos for Person Tracking*, 106 (2-3), May - June 2007.

Registration of Multimodal Imagery with Occluding Objects using Mutual Information
S. J. Krotosky and M. M. Trivedi, *Applied Perception in Thermal Infrared Imagery*, In Press.

Registration of Multimodal Stereo Images using Disparity Voting from Correspondence Windows
S. J. Krotosky and M. M. Trivedi, *Proc. IEEE International Conference on Advanced Video and Signal based Surveillance*, November 2006.

Multimodal Stereo Image Registration for Pedestrian Detection
S. J. Krotosky and M. M. Trivedi, *Proc. IEEE Conference on Intelligent Transportation Systems*, September 2006.

Multimodal Image Registration for Person Detection: Analysis and Review
S. J. Krotosky, *CVRR Technical Report*, December 2005.

Real-Time Stereo-Based Head Detection using Size, Shape and Disparity Constraints
S. J. Krotosky, S. Y. Cheng and M. M. Trivedi, *Proc. IEEE Intelligent Vehicles Symposium*, June 2005.

Face Detection and Head Tracking using Stereo and Thermal Infrared Cameras for "Smart" Airbags: A Comparative Analysis
S. J. Krotosky, S. Y. Cheng and M. M. Trivedi, *Proc. IEEE Conference on Intelligent Transportation Systems*, October 2004.

Occupant Posture Analysis with Stereo and Thermal Infrared Video: Algorithms and Experimental Evaluation
M. M. Trivedi, S. Y. Cheng, E. M. C. Childers and S. J. Krotosky, *IEEE Transactions on Vehicular Technology, Special Issue on In-Vehicle Vision Systems*, 56 (6), November 2004.

Occupant Posture Analysis using Reflectance and Stereo Images for "Smart" Airbag Deployment
S. J. Krotosky and M. M. Trivedi, *Proc. IEEE Intelligent Vehicles Symposium*, June 2004.

Detection and identification of sardine eggs at sea using a machine vision system
J. R. Powell, S. J. Krotosky, B. Ochoa, D. Checkley and P. Cosman, *Proc. OCEANS*, 2003.

ABSTRACT OF THE DISSERTATION

On Using Multiperspective Color and Thermal Infrared Videos to Detect People: Issues, Computational Framework, Algorithms and Comparative Analysis

by

Stephen Justin Krotosky
Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)
University of California, San Diego, 2007
Professor Mohan M. Trivedi, Chair

This is a study to investigate the fundamental problem of combining color and infrared imagery in a unified feature framework that can then be applied to person detection. In order to combine the imagery, the features of objects in the scene must be registered. This is a challenge in color and infrared imagery, as corresponding features appear very different in each image spectrum. Once registered it is also a challenge to successfully combine the features to achieve improved detection over unimodal approaches. We investigate both these challenges in detail.

We present the related studies in multimodal image registration and categorize the registration methodologies into four distinct sectors based on the assumptions about scene configuration. We examine how these assumptions limit the generality of scenes that can be analyzed and help motivate the development of an approach to registering color and infrared imagery that is able to overcome these limitations.

In order to register multiple objects in a general scene, where objects can be at different depths from the camera, stereo analysis is necessary to resolve the parallax associated with the multiple views. We first examine state-of-the-art stereo algorithms that are designed to handle correspondence matching for unmatched image data. We definitively show that these approaches are unsuitable for finding correspondence in cross-spectral stereo imagery, where a color and infrared camera are joined in a stereo

pair. As an alternative, we propose a region-based approach to correspondence matching that is able to successfully perform correspondence matching by relying on an initial segmentation and disparity voting-based methodology to registering foreground objects in the scene.

Extensive experimental evaluations of our proposed cross-spectral stereo registration algorithm are performed. We present experimental studies in registering people in both indoor surveillance from a static camera and outdoor pedestrian detection from a moving vehicle. We also offer a comparison of our approach to ground truth and the current state of related studies, with both ideal and realistic initial segmentations. We also experimentally validate the robustness of our approach by evaluating additional data taken from different cameras in another environment. Finally, we show how our approach to cross-spectral stereo registration can be used to track people in a 3D context.

Our study then focuses on studying how color and infrared imagery can be used to improve person detection algorithms. In the context of pedestrian detection, we first compare and evaluate how the disparity information from color stereo and infrared stereo can be used to detect potential objects in the scene. The high success of the disparity information from both modalities motivates a discussion of the color and infrared features that can be extracted to further classify the potential objects into pedestrian and non-pedestrian regions. This leads to our development of our experimental framework that allows us to compare pedestrian classifiers that utilize all combinations of color, infrared and disparity features. We also propose a trifocal framework consisting of a color stereo camera rig combined with an infrared camera in order to quickly register the multimodal data for our analysis.

We extend the analysis of multispectral and multiperspective approaches to person detection in the context of surveillance. We further justify our trifocal approach to registration by demonstrating its superiority over the planar homography approach in terms of scene generality and robustness. The trifocal approach is able to register any object in the scene that is able to be registered in stereo imagery. This allows general scene configurations and also allows for a direct comparison to conventional monocular

and unimodal stereo approaches. With this in mind, we present a framework for person detection that can combine color, infrared and disparity features in a unified manner and expands the robustness and accuracy of the method proposed in the previous chapter. We then use this algorithmic framework to present a detailed comparison of person detection using various combinations of color, infrared and disparity features. The analysis demonstrated that our unified trifocal framework easily outperforms both unimodal stereo analysis and multimodal "tetravision" analysis that separately combines color and infrared stereo analysis. We present extensive evaluation of the trifocal-based experiments to illustrate the improved detection rates that can be achieved when incorporating multispectral data in the detection framework.

Chapter I

Introduction

I.A Motivation

The analysis of people is a general topic of interest in many fields. Specific interest lies in determining how people act and interact in "intelligent" environments. Such spaces are equipped with sensors that can derive and maintain an awareness of the events and activities that occur in the space. These types of spaces can include indoor environments and buildings, as well as outdoor spaces, moving vehicles and any other spaces that humans occupy.

Cameras and video networks make natural sensor systems for "looking" at people. Intelligent environments need to support a wide and general set of person actions and interactions; video analysis and computer vision techniques provide a natural framework for this. When analyzing people it is desirable to have a system that can provide multi-level descriptions of the human activity, including tracking people in 3-D, estimating their poses and identifying their interactions with the environment and others. Each of these goals can be addressed with computer vision techniques.

When looking at people in this context, it is desirable to obtain as much information as possible to aid in accurately and robustly analyzing the scene. With this in mind, a multi-perspective approach can be used when imaging the intelligent environment. By viewing the scene from multiple perspectives, 3-D information can be extracted from the scene making it easy to detect, track and analyze people in the scene. Additionally, it is important to be able to distinguish people in the scene. There are many techniques for analyzing people in color imagery, but those techniques, and color imagery in general, are susceptible to lighting conditions. In order to provide robustness to this, we would like to combine our color camera analysis with thermal imagery. The thermal imagery will add to the robustness of the system by providing an additional way of viewing the scene.

It is within this multimodal and multi-perspective framework that we wish to explore person analysis.

I.B Outline

In Chapter II, related studies in multimodal image registration are reviewed. We categorize the registration methodologies into four distinct sectors based on the assumptions about scene configuration. We examine how these assumptions limit the generality of scenes that can be analyzed and help motivate the development of an approach to registering color and infrared imagery that is able to overcome these limitations.

Chapter III details our approach to multimodal image registration. We show that in order to register multiple objects in a general scene, where objects can be at different depths from the camera, stereo analysis is necessary to resolve the parallax associated with the multiple views. We first examine state-of-the-art stereo algorithms that are designed to handle correspondence matching for unmatched image data. We definitively show that these approaches are unsuitable for finding correspondence in cross-spectral stereo imagery, where a color and infrared camera are joined in a stereo pair. As an alternative, we propose a region-based approach to correspondence matching that is able to successfully perform correspondence matching by relying on an initial segmentation and disparity voting-based methodology to registering foreground objects in the scene.

In Chapter IV, we perform an extensive experimental evaluation of our proposed cross-spectral stereo registration algorithm. We present experimental studies in registering people in both indoor surveillance from a static camera and outdoor pedestrian detection from a moving vehicle. We also offer a comparison of our approach to ground truth and the current state of related studies, with both ideal and realistic initial segmentations. We also experimentally validate the robustness of our approach by evaluating additional data taken from different cameras in another environment. Finally, we show how our approach to cross-spectral stereo registration can be used to track people in a 3D context.

The next two chapters shifts focus from the problem of registering color and infrared imagery to studying how color and infrared imagery can be used to improve

person detection algorithms. In Chapter V explore using color and infrared imagery for detecting people in the context of pedestrian detection. We first compare and evaluate how the disparity information from color stereo and infrared stereo can be used to detect potential objects in the scene. The high success of the disparity information from both modalities motivates a discussion of the color and infrared features that can be extracted to further classify the potential objects into pedestrian and non-pedestrian regions. This leads to our development of our experimental framework that allows us to compare pedestrian classifiers that utilize all combinations of color, infrared and disparity features. We also propose a trifocal framework consisting of a color stereo camera rig combined with an infrared camera in order to quickly register the multimodal data for our analysis. We will explore this trifocal framework and classification architecture much further in Chapter VI.

Chapter VI continues our analysis of multispectral and multiperspective approaches to person detection in the context of surveillance. We further justify our trifocal approach to registration by demonstrating its superiority overthe planar homography approach in terms of scene generality and robustness. The trifocal approach is able to register any object in the scene that is able to be registered in stereo imagery. This allows general scene configurations and also allows for a direct comparison to conventional monocular and unimodal stereo approaches. With this in mind, we present a framework for person detection that can combine color, infrared and disparity features in a unified manner and expands the robustness and accuracy of the method proposed in the previous chapter. We then use this algorithmic framework to present a detailed comparison of person detection using various combinations of color, infrared and disparity features. The analysis demonstrated that our unified trifocal framework easily outperforms both unimodal stereo analysis and multimodal "tetravision" analysis that separately combines color and infrared stereo analysis. We present extensive evaluation of the trifocal-based experiments to illustrate the improved detection rates that can be achieved when incorporating multispectral data in the detection framework.

Chapter VII summarizes the work and presents the concluding remarks.

Chapter II

Related Studies in Multiperspective Image Registration

II.A Introduction

A fundamental issue associated with multisensory vision is that of accurately registering corresponding information and features from the different sensory systems. This issue is exacerbated when the sensors are capturing signals derived from totally different physical phenomena, such as color (reflected energy) and thermal signature (emitted energy). Multimodal imagery applications for human analysis span a variety of application domains, including medical [1], in-vehicle safety systems [2] and long-range surveillance [3]. The combination of both types of imagery yields information about the scene that is rich in color, depth, motion and thermal detail. Once registered, such information can then be used to successfully detect, track and analyze movement and activity patterns of persons and objects in the scene.

At the heart of any registration approach is the selection of the most relevant similarity metric, which can accurately match the disparate physical properties manifested in images recorded by multimodal cameras. Mutual Information (MI) provides an attractive metric for situations where there are complex mappings of the pixel intensities of corresponding objects in each modality, due to the disparate physical mechanisms that give rise to the multimodal imagery [4]. Egnal has shown that mutual information is a viable similarity metric for multimodal stereo registration when the mutual information window sizes are large enough to sufficiently populate the joint probability histogram of the mutual information computation [5]. Further investigations into the properties and applicability of mutual information for windowed correspondence measure has been done by Thevenaz and Unser [6]. Challenges lie in obtaining these appropriately sized window regions for computing mutual information in scenes with multiple people and occlusions, where a balanced tradeoff between larger windows for matching evidence and smaller windows for registration detail is needed.

In this chapter, we provide a detailed overview of the current state of multimodal registration and provide a comparative analysis of algorithms for registering color and infrared image information. A discussion of the pros and cons of each ap-

proach helps motivate the development of the cross-spectral stereo approach we will discuss in subsequent chapters.

II.B Multimodal Registration Approaches: Comparative Analysis of Algorithms

In a multimodal, multicamera setup, because each camera can be at a different position in the world and have different intrinsic parameters, objects in the scene can not be assumed to be located at the same position in each image. Due to these camera effects, corresponding objects in each image may have different sizes, shapes, positions, and intensities. In order to combine the information in each image, it is required that the corresponding objects in the scene be aligned, or registered. Sensory measurements can then be fused or features combined in a variety of ways that can fuel algorithms that take advantage of the information provided from multiple and differing image sources [7]. Experiments in our previous work [2] have offered analysis and insight into the commonalities and uniqueness of the multimodal imagery. Multimodal image registration approaches vary based on factors such as camera placement, scene complexity and the desired range and density of registered objects in the scene. In order to better understand the algorithmic details of the various multimodal registration techniques, it is important to outline the underlying geometric framework for registration. Much of the multiple view geometry properties derived in this paper are adapted from Hartley and Zisserman [8].

Given a two camera setup with camera center locations C and C', a 3D point in space can be defined relative to each of the camera coordinate systems as $P = (X, Y, Z)^T$ and $P' = (X', Y', Z')^T$, respectively. The coordinate system transformation between P and P' is:

$$P' = RP + T \tag{II.1}$$

where R is the matrix that defines the rotation between the two camera centers and

T is the translation vector that represents the distance between them. Additionally, the projection matrices for each camera are defined as K and K', where the projected points on the image plane are the homogeneous coordinates $p = (x, y, 1)$ and $p' = (x', y', 1)$.

Let π be a plane in the scene parameterized with N, the surface normal of the plane and d_π is the distance from the camera center C. Then a point lies on that plane if $N^T P = d_\pi$. The homography induced by π is $P' = H_P P$ where:

$$H_P = R - T\,\frac{N^T}{d_\pi} \tag{II.2}$$

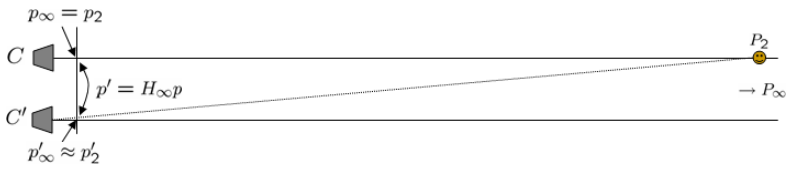
Applying the projection matrices K and K', we have $p' = Hp$, where $H = K' H_P K^{-1}$ giving

$$H = K'\big(R - T\,\frac{N^T}{d_\pi}\big)K^{-1} \tag{II.3}$$

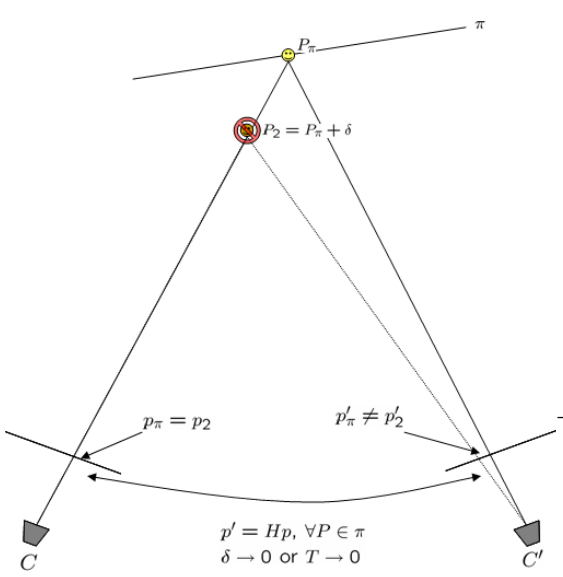
This homographic transformation describes the transformation of points only when the points lie on the plane π (e.g. $N^T P = d_\pi$). When a point does not lie on this plane, then an additional parallax component needs to be added to the transformation equation to accommodate the projective depth of other points in the scene relative to the plane π. It has been shown in [8] that the transformation that includes the additional parallax term is:

$$p' = Hp + \delta e' \tag{II.4}$$

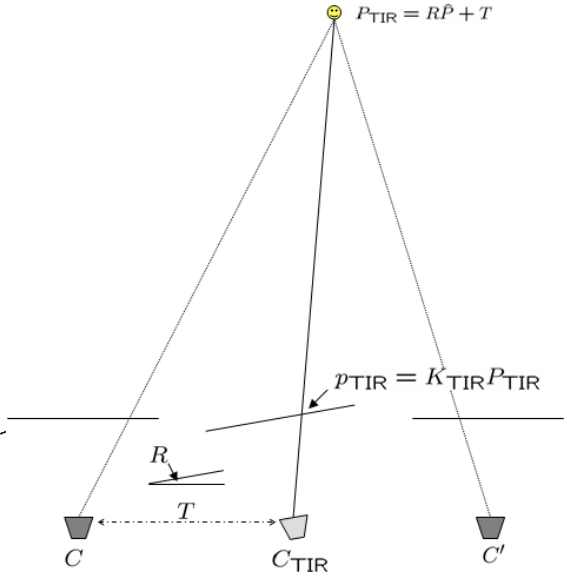
where e' is the epipole in C' and δ is the parallax relative to the plane π. The epipole is the intersecting point between the image plane and the line containing the optical centers of C and C'. The equation in (II.4) has effectively decomposed the point correlation equation into a term for the induced planar homography (Hp) and the parallax associated with points that do not satisfy the planar homography assumption ($\delta e'$). It is within this framework that we will describe the registration techniques used for multimodal imagery. Figure II.1 illustrates the main approaches to multimodal image registration that will be analyzed. Additionally, Table II.1 provides a summary of references utilizing these approaches and indicates the assumptions, methods and limitations in each.



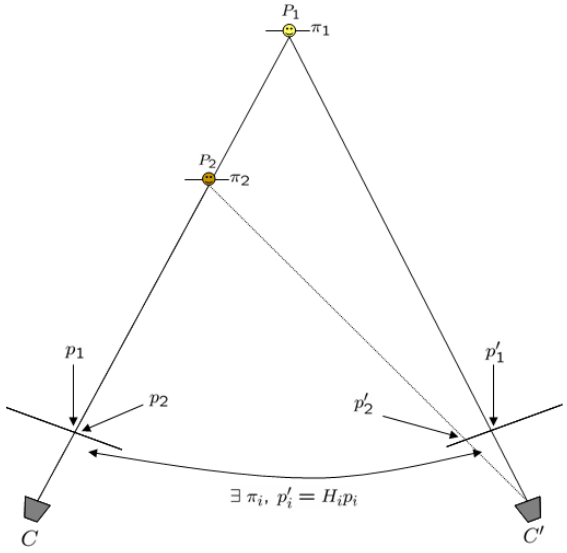
(a) Infinite Homography



(b) Global



(c) Stereo Geometric



(d) Partial Image ROI

Figure II.1: Geometric illustration of the four main approaches to multimodal image registration.

II.B.1 Infinite Homographic Registration

In O'Conaire *et al.* [9] and Davis & Sharma [3], it is assumed that the thermal infrared and color cameras are nearly colocated and the imaged scene is far from the camera, so that the deviation of pedestrians from the ground plane is negligible compared to the distance between the ground and the cameras. Under these assumptions, an infinite planar homography can be applied to the scene and all objects will be aligned in each image.

The infinite planar homography, H_∞, is defined as the homography that occurs when the plane π is at infinity. An illustration of this type of registration geometry is shown in Figure II.1(a). Starting from (II.3), we define

$$H_\infty = \lim_{d_\pi \to \infty} H = K' R K^{-1} \tag{II.5}$$

When the plane is at infinity, the homography between points is only a rotation R between the cameras and the internal projection matrices for each camera, K and K'. Similarly, from (II.4), Hartley and Zisserman [8] showed that the correspondence equation for image points in an infinite homography is:

$$p' = H_\infty p + \frac{K't}{Z} \tag{II.6}$$

where $Z = \frac{1}{\delta}$ is the depth from C and $K't = e'$ is the epipole in C'.

Infinite homographic registration techniques are used when the scene distance is very far from the camera. When all observed objects are very far from C, then $Z \to \infty$ and the parallax effects will be negligible. Alternatively, when the cameras are nearly colocated, i.e. $t \to 0$, the parallax term also becomes negligible. In both cases the correspondence equation becomes:

$$p' = H_\infty p \tag{II.7}$$

The use of an infinite planar homography is a an effective way of registering the scene, but only when the scene that is being registered conforms to the homographic

assumptions. This means that the scene must be very far from the camera so that an object's displacement from the ground plane will be negligible compared to the observation distance. While this type of assumption is appropriate for long distance and overhead surveillance scenes, this is not valid in situations where objects and people can be at various depths whose difference is significant relative to their distance from the camera. In these cases, the infinite homography assumption will not align all objects in the scene. In addition, when the assumption of an infinite homography does hold, the lack of a parallax term precludes any estimate of depth that could be used as a differentiator for occluding objects.

II.B.2 Global Image Registration

Global approaches to registration can be used when further assumptions about the movement and placement of objects and people in a scene are employed to make the registration fit a specific model. The registration will be accurate when the scene follows the specific model used, but can be grossly inaccurate when the imaged scene does not fit the assumptions of the model.

The usual assumption of these techniques is that all objects lie on the same plane in the scene. Often to enforce this assumption, only foreground objects are considered. Global image registration techniques make the assumption that δ, the measure of difference from the homographic plane in (II.4), will be small for all objects in the scene. However, in scenes where objects of interest are at different planes, only the objects lying on the plane π that induces the homography will be registered. All other objects that lie on different planes will be misaligned due to the second term $\delta e'$ in (II.4).

If the distance of objects from the plane is small compared to the distance of cameras from the plane, the parallax effects tend to zero and the homography accurately describes the registration of objects in the scene at any depth. Works that have applied this global registration technique operated either on the single plane or approximate colocation assumption to allow for accurate scene registration. An illustration of this type of registration is shown in Figure II.1(b).

Irani and Anandan [10] used directional-derivative-energy operators to generate features from a Gaussian Pyramid of the visual and thermal infrared images and used local correlation values for these features to obtain a global alignment for the multimodal image pair. Alignment is done by estimating a parametric surface correspondence that can estimate the registration alignment of the two images. Newton's method is used to iteratively search for the parametric transformation that maximizes the global alignment.

Coiras *et al.* [11] matches triangles formed from edge features in visual and thermal infrared images to learn an affine transformation model for static images. The global affine transformation that best maximizes the global edge-formed triangle matching is searched from transformations obtained by matching individual formed triangles in one image to other individual formed triangles in the second image.

Han and Bhanu [12] used the features extracted when a human walked in a scene to learn a projective transformation model to register visual and IR images. It is assumed that the person walking in the scene walks in a straight line during the registration sequence. This enforces that the person is located within a single plane throughout the sequence and ensures that the global projective transformation model assumption holds. Feature points derived from foreground silhouettes in two pair of images in the sequence are used as input into a Hierarchical Genetic Algorithm that searches for the best global transformation.

Itoh *et al.* [13] used a calibration board to register colocated color and thermal infrared cameras for use in a system that recognized hand movement for multimedia production. The calibration board points were used to establish a quadratic transformation model between the color and thermal infrared images. Registration is only required for a predefined workspace with a fixed range within the image scene and the calibration board was placed to ensure registration in that region.

Similarly, Ye [14] used silhouette tracking and Hausdorff distance edge matching to register visual and thermal infrared images. In this case, it is assumed that the cameras are nearly colocated and that registration can be accomplished with a displacement and scaling. The detected top points of foreground silhouettes are tracked using

the motion associations with previously tracked points. The Hausdorff distance measure is used to match edge features in each silhouette and estimate the scale and translation parameters. The registration and tracking are then used and updated to provide simultaneous tracking and iterative registration.

Global image registration methods place some limiting assumptions on the configuration of objects in the scene. Specifically, it is assumed that all registered objects will lie on a single plane in the image and it is impossible to accurately register objects at different observation depths, as the registration transform for each object will depend on the varying perspective effects of the camera. This means that accurate registration can only occur when there is only one observed object in the scene [12], or when all the observed objects are restricted to lie at approximately the same distance from the camera [13] [14]. The global alignment algorithms proposed by Irani & Anandan [10] and Coiras *et al.* [11] do not account for situations where there are objects at different depths or planes in the image. Both use the assumption that the colocation of the cameras and the observed distances are such that the parallax effects can be ignored.

The primary limitation to global registration methods is that it is impossible to register objects at different depths. Global methods effectively restrict the successfully registered area to be a single plane in the image. When colocated cameras are used to relax the single plane restriction, parallax effects become negligible, and the problem becomes akin to infinite homographic methods.

II.B.3 Stereo Geometric Registration

When a stereo camera setup is used in combination with additional cameras from other modalities, the images from each modality can be combined using the stereo 3D point estimates and the geometric relation between the stereo and multimodal cameras. As demonstrated in Ju *et al.* [15], stereo cameras can give accurate 3D point coordinates for objects in the image. If the remaining cameras are then calibrated to the reference stereo pair, usually with a calibration board, then the pixels in those images (thermal infrared) can be reprojected onto the reference stereo image. The resulting

reprojection will be registered to the stereo reference image.

In this case, for a point p in the reference stereo image, an estimate of its 3D location $\hat{P}$ is given from the calibrated stereo geometry parameters. Additionally, the calibration of the left reference stereo image and the additional thermal infrared modality give the rotation R and T between camera coordinates. This allows the change of coordinate system to the thermal infrared reference frame, $P_{\text{TIR}} = R\hat{P} + T$. The 3D point can then be reprojected onto the infrared image plane.

$$p_{\text{TIR}} = K_{\text{TIR}} P_{\text{TIR}} \tag{II.8}$$

The thermal image point is then put into homogeneous form and the intensity value at this location in the thermal infrared image can then be assigned to the point p in the stereo reference image. Such a registration technique is illustrated in Figure II.1(c).

For the case of stereo geometric registration techniques, objects in a scene at very different depths can be registered as long as the stereo disparity information is available for that object. If the stereo algorithm can provide dense and accurate stereo for the objects in the scene, stereo geometric registration is a good way of quickly and effectively registering the visual and infrared imagery. In the experiments of Ju *et al.* [15] the observed object (head) was carefully placed into the scene and it was assumed that it was the only object in the scene. Stereo data was captured using high resolution stereo cameras in a fairly stable and well-conditioned scene. The resulting 3D stereo image was dense and accurate in these conditions. However, experiments need to be conducted to see how these environmental conditions can be relaxed. Namely, it is important to examine how stereo geometric registration techniques perform in *real world* conditions, where using standard resolution cameras in environments of poor lighting, poor textures and occlusions can affect the quality and reliability of the 3D reprojection registration technique.

Multiple stereo camera approaches to stereo geometric have been investigated by Bertozzi *et al.* [16]. Using four cameras configured into two unimodal stereo pairs that yield two separate disparity estimates, registration can occur in the disparity domain.

While this approach yields redundancy and registration success, the use of four cameras can be cumbersome both in physical creation, calibration and management, as well as in data storage and processing.

II.B.4 Partial Image ROI Registration

An approach to registering objects at multiple depths is to use partial image region-of-interest registration. The main assumption of this approach is that each individual object in the scene is at a specific plane and that each plane can be individually registered with a separate homography. For each of the i regions-of-interest Ω in the image, if $p \in \Omega_i$ then

$$p' = H_i p + \delta_i e' \tag{II.9}$$

Again, it assumed that the parallax effects are negligible within each object, as each is approximated to be a single planar object in the scene. As long as each Ω_i satisfies this assumption, the registration technique will be applicable. This is illustrated in Figure II.1(d).

Chen *et al.* [17] proposed that the visual and infrared imagery be registered using a maximization of mutual information technique on bounding boxes that correspond to detected objects in one of the modalities. It is assumed that corresponding regions can be found by translation. It is also assumed that any scale difference is fixed and known a priori. The matching bounding box is then searched for in the other modality using a simplex method. This allows bounding boxes that correspond to objects at different depths to be successfully registered.

Chen *et al.* assume that the bounding boxes representing a single object can always be properly segmented and tracked in one of the modalities. The assumption that bounding boxes will be properly segmented will often not hold, especially in uncontrolled scenes where the issues of lighting, texture and occlusions can produce segmentation results that contain two or more merged objects at different depths. Using bounding boxes that contain multiple objects will not register properly as the required assumption that an ROI contains objects within a single plane will not hold.

Table II.1: Review of Approaches to Multimodal Registration and Body Analysis

Work	Modalities Vis	IR	3D	Calib.	Registration	Comments
Trivedi *et al.*(2004) [2]	X	X	X		None	Comparative Analysis of Head Detection Algorithms using both stereo and thermal infrared imagery.
Davis & Sharma (2004, 2005) [18] [19] [3]	X	X			Infinite Homography	Requires observed scene far from colocated cameras.
O'Conaire *et al.*(2005) [9]	X	X			Infinite Homography	Requires observed scene far from colocated cameras.
Irani & Anandan (1998) [10]	X	X			Global	Experimental images only contain one dominant plane in scene and no foreground objects. Global parametric model not likely to model large parallax effects well.
Coiras *et al.*(2000) [11]	X	X			Global	Global affine model cannot account for large parallax effects. Experiments are not performed for multiple objects in scene at different planes.
Han & Bhanu (2003) [12]	X	X			Global	Walking along different planes results in different registration. Multiple people at different depths will not be registered. Unrealistic that humans walk in same line for registration. Need entire sequence before first frame can be registered.

TableTable II.1 – continued from previous page

Work	Modalities Vis	IR	3D	Calib.	Registration	Comments
Itoh *et al.*(2003) [13]	X	X	X	X	Global	Registration assumptions only valid for objects within a range of certain depths located inside the limited "workspace". Information from each modality is heuristically thresholded and not probabilistically generalized.
Ye (2005) [14]	X	X			Global	Global matching not valid when people are at large depth differences. Experiments do not test large movements over sequences where registration parameters would be changing.
Ju *et al.*(2004) [15]	X	X	X	X	Stereo Geometric	Registration evaluation needed for in low-res stereo environment and in real-world conditions, e.g. multiple people, occlusions, lighting, etc.
Bertozzi *et al.*(2006) [16]	X	X	X	X	Stereo Geometric	Four camera system cumbersome in terms of setup and maintenance, as well as in terms of image processing and data management. Color and Infrared registered only at bounding box level.
Chen,*et al.*(2003) [17]	X	X			Partial Image ROI	Assumption of perfect target tracking gives ideal bounding boxes. Need to handle occlusions, overlaps, and incompleteness.
Our Approach [20]	X	X	X	X	Disparity Voting	Successful registration through occlusions and scenes with multiple people. Disparity estimates can be used as feature in tracking algorithms.

II.C Summary

In this chapter we have provided an analysis of the approaches to multimodal image registration and detailed the assumptions, applicability and limitations of each. We have shown how current approaches restrict registration to scenes that fall under specific configurations that severely limit the general applicability of multimodal analysis. To generalize the registration to include objects that are different depths from the camera, multiperspective elements must be used to account for the parallax in the scene. It is in this multiperspective and multimodal domain that we will focus our efforts in the subsequent chapters.

The text of this chapter, in part, is a reprint of the material as it appears in: Stephen J. Krotosky and Mohan M. Trivedi, "Mutual Information Based Registration of Multimodal Stereo Videos for Person Tracking" in *Computer Vision and Image Understanding, Special Issue on Advances in Vision Algorithms and Systems Beyond the Visible Spectrum*, Vol. 106, Issues 2-3, May-June 2007. I was the primary researcher of the cited material and the co-author listed in this publication directed and supervised the research which forms a basis for this chapter.

Chapter III

Cross-Spectral Stereo

III.A Introduction

A fundamental issue associated with stereo vision is finding accurate and robust similarity measures that can be used to match correspondences across different sensory systems. This issue is exacerbated when the capture devices yield image intensities that are derived from totally different physical phenomena, such as color and infrared imagery. While a multimodal stereo system has the potential to give information about the scene that is rich in color, depth, motion and thermal detail, it is necessary to first find similarity measures that resolve the multimodal stereo correspondences.

One such similarity measure, mutual information (MI), provides an attractive metric for situations where there are complex mappings of the pixel intensities of corresponding objects in each modality. Egnal has shown that mutual information is a viable similarity metric for multimodal stereo registration when the mutual information window sizes are large enough to sufficiently populate the joint probability histogram of the mutual information computation [5]. Further investigations into the properties and applicability of mutual information for windowed correspondence measure has been done by Thevenaz and Unser [6]. Challenges for multimodal stereo imagery lie in appropriately applying mutual information similarity measures in a way that can efficiently resolve the correspondence problem. We will investigate several methods for achieving this, namely an energy minimization framework and a region and edge segment-based multiprimitive framework.

This chapter presents the following contributions: In Section III.B, we give a detailed analysis of current approaches to stereo matching that use mutual information in an energy minimization framework. We experimentally demonstrate that these methods cannot resolve the stereo correspondences when using true multimodal imagery. In Section III.C we present our alternative approach to resolving the stereo correspondences by focusing on matching region-based primitives. This approach is able to successfully register multiple objects in the scene at significantly different depths from the camera.

III.B Cross-Spectral Stereo using Mutual Information

Recently, algorithms have been developed that utilize mutual information to solve the stereo correspondence between two images. Using mutual information to measure the similarity of potential correspondences is attractive because it is inherently robust to differences in intensities between two corresponding points. Egnal [5] is historically attributed with proposing the idea of using mutual information as a stereo correspondence matching feature, yet results were of relatively low quality until Kim *et al.* [21] and subsequently Hirschmüller [22] demonstrated very successful stereo disparity generation by using mutual information in an energy minimization context. They have shown how the mutual information measure gives good results even when the images are synthetically altered by an arbitrary intensity transformation. We investigate whether these mutual information based stereo algorithms can resolve the correspondence problem for true multimodal imagery with the same success achieved for synthetically altered imagery.

We have chosen to utilize the algorithm developed by Hirschmüller [22] in analyzing the use of mutual information with energy minimization for solving multimodal stereo correspondences. This choice is based on the fact that this algorithm is the mutual information-based approach that performed best on the Middlebury College Stereo Evaluation [23]. Its use of mutual information is identical to that of Kim *et al.* [21] and the two algorithms differ only in how the energy function is minimized, with Kim using the global optimization of *Graph Cuts* while Hirschmüller utilizes a faster hierarchical approach called *Semi-Global Matching*.

To compute mutual information in this framework, Kim *et al.* adapted the mutual information computation to fit within the energy minimization framework. We re-derive this computational framework here for convenience. The mutual information (MI) between two images I_L and I_R is defined as:

$$MI_{L,R} = H_L + H_R - H_{L,R} \tag{III.1}$$

where H_L and H_R are the entropies of the two images and $H_{L,R}$ is the joint entropy term. These entropies are defined as:

$$H_L = -\int P_L(l) \log P_L(l) dl \tag{III.2}$$

$$H_{L,R} = -\int\int P_{L,R}(l,r) \log P_{L,R}(l,r) dl dr \tag{III.3}$$

where P is the probability distribution of intensities for a given image (L) or image pair (L,R), respectively. In order to put the entropy terms into the energy minimization framework, Kim approximated the H as a sum of terms based on each pixel pair p in the imagery:

$$H_{L,R} = \sum_p h_{L,R}(L_p, R_p) \tag{III.4}$$

The joint entropy, $h_{L,R}$ is computed performing Parzen estimation (2D convolution with Gaussian $g(l,r)$) and approximating the probability distribution $P_{L,R}$ as the normalized 2D histogram of corresponding pixels from image pair I_L and I_R.

$$h_{L,R} = -\frac{1}{n} \log(P_{L,R}(l,r) \otimes g(l,r)) \otimes g(l,r) \tag{III.5}$$

Similarly, the entropy term is:

$$h_L = -\frac{1}{n} \log(P_L(l) \otimes g(l)) \otimes g(l) \tag{III.6}$$

From this, Kim redefined mutual information as:

$$MI_{L,R} = \sum_p mi_{L,R}(L_p, R_p) \tag{III.7}$$

$$mi_{L,R}(l,r) = h_L(l) + h_R(r) - h_{L,R}(l,r) \tag{III.8}$$

It is this mi term that both Kim [21] and Hirschmüller [22] use in their iterative stereo algorithm cost functions. We experiment with the stereo algorithm proposed by Hirschmüller for a variety of multimodal imagery, including color pairs, synthetically altered color pairs and paired color/infrared imagery.

(a) Color Imagery

(b) Posterized Color Imagery

(c) Synthetically Altered Color Imagery

(d) Infrared Imagery

(e) Color and Infrared Imagery

Figure III.1: Mutual Information Stereo Examples: Disparity results from Mutual Information based stereo algorithm for different input images. Notice how disparity values are reasonable even for highly altered inputs, but the algorithm fails for natural multimodal image sets.

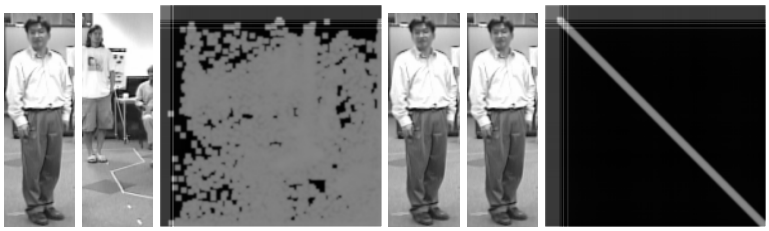
Figure III.1 shows the results of the Semi-Global Matching algorithm using mutual information proposed in [22] for different test images. The first row shows the results for two matched color stereo pairs. Notice how the resulting disparity image provides dense and quality estimates for the entire image. For each object in the scene, there is a silhouette of disparity that fits logically with the scene. Depth order is maintained throughout and the overall disparity image appears similar to those reported in the stereo matching literature [23]. These results are expected and are on par with the quality of disparity results reported in the original paper. The results in the second row show the disparity image when the right image is posterized to 8 intensity levels. The results in the third row show when the right image is synthetically altered with an arbitrary transform. In this case, the transform is quite complex and the intensity transform is not one-to-one, $y = 128(cos(x/15) \cdot x/255 + 1)$. Each of these disparity images gives dense and accurate estimates that are very similar to the original unaltered stereo pair. This assessment corroborates with other stereo results for synthetically altered imagery reported in [21] and [22]. Additionally, the fourth row shows successful stereo matching when using two infrared images.

The final row, Figure III.1(e), shows the results when the same algorithm is applied to multimodal stereo imagery. The resulting disparity image yields completely invalid results and the algorithm cannot resolve any of the correct correspondences. The people in the infrared image are clearly visible and we as humans would have no problem finding the corresponding person from the color image. The transform between color and thermal, while different from the synthetic transform, does not appear to be markedly worse, although some details, especially in the background regions, are lost. The question remains, what is fundamentally different about the infrared imagery that prevents the correct determination of correspondence values?

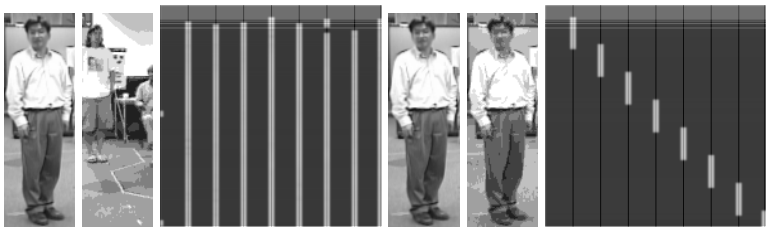
To try to answer this question, we need a deeper analysis of the underlying mutual information optimization scheme. At the initialization of the energy minimization algorithms, a random disparity map is chosen to initialize the probability distribution that is used to compute the mutual information terms. At this point, it is expected that

the mutual information, denoted mi in [22] and D in [21], will appear relatively uncorrelated and give a low mutual information score. As the algorithmic iterations progress, it is desired that the mi values approach a maximum and the 2D mi plot follows the true intensity relation between the left and right images. For example, for the matched color stereo pair, the mi values lie along a line with negative unit slope when the correct disparity correspondences are found.

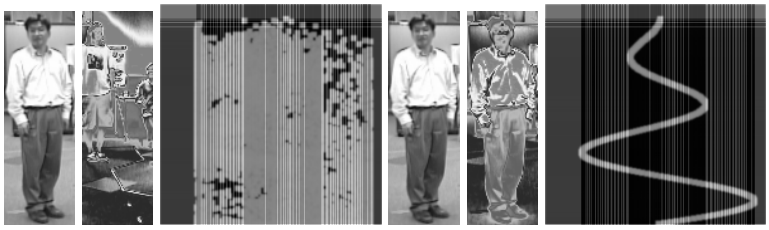
Figure III.2 shows the mi plot for a pair of non-corresponding and corresponding regions for color-color (a), color-posterized (b), color-altered color (c) infrared-infrared (d) and color-infrared (e) imagery. The first pair of images of each row can be thought of as starting from an initially random disparity image where most (or all) of the correspondences are incorrect. In this case, the resulting mi plot shows intensities that are not well correlated as noted by its large spread across the image 2D mi histogram. For the color-color, color-posterized, color-altered color and infrared-infrared cases, when we choose corresponding image regions, the mi plot shows the well correlated image intensity transform, as expected. However, for the case of corresponding color-infrared images, the mi value does not reduce to some easily discernable transform. In fact, the intensities for the corresponding multimodal regions appear just as uncorrelated as the intensities for the non-corresponding regions. This indicates that using these types of energy minimization algorithms is not possible with color and infrared stereo imagery. This uncorrelatedness of the color and thermal imagery means that it is difficult to predict the intensity of an infrared pixel given a corresponding color intensity. Because of this, the use of mutual information as an energy minimization term is not appropriate. The mutual information energy term (mi values) needs to be minimized, yet cannot because the uncorrelation between color and thermal image intensities produces similarly large values for both good and bad matches.



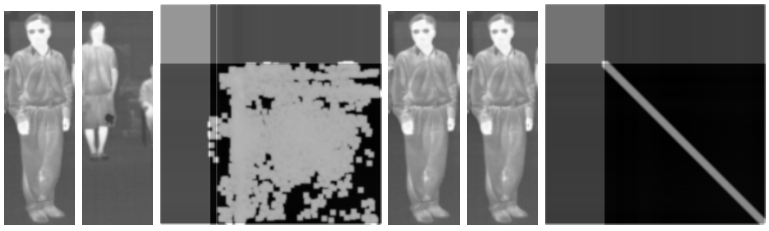
(a) Color-Color MI Transform



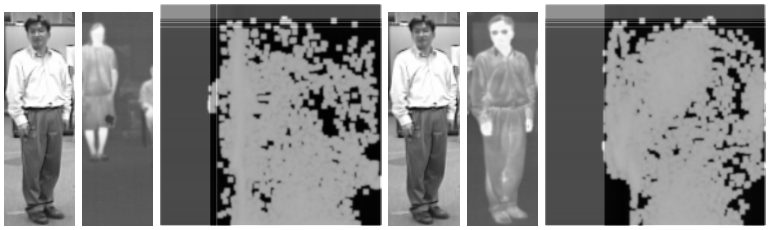
(b) Color-Posterized Color MI Transform



(c) Color-Altered Color MI Transform



(d) Infrared-Infrared MI Transform



(e) Color-Infrared MI Transform

Figure III.2: mi plots for non-corresponding and corresponding image regions.

III.C Multimodal stereo using primitive matching

We have demonstrated that current state-of-the-art stereo algorithms cannot utilize mutual information to effectively solve the multimodal stereo correspondence problem. It is important to now seek out alternative features and approaches that may give some way of obtaining correspondences in the scene. To achieve any success in stereo correspondence matching with multimodal imagery, it is imperative to first identify features that are universal to both the color and thermal imagery. While it is clear that there is little commonality associated with the intensities across color and thermal imagery, the example multimodal stereo pair in Figure III.4 suggests that there is some clear commonality on a region (object) level and on edges associated with these region boundaries. For example, skin tone regions in the color image correspond well to bright intensity regions on the infrared image. In general, the silhouettes associated with the people in the scene have similar sizes, shapes and edge boundaries in each modality.

Prior to the success and proliferation of windowed correlation-based approaches to the traditional stereo correspondence problem, many researchers attempted to solve the correspondences between two images by utilizing region based primitives. As it is apparent that regions share strong commonality across the multimodal imagery, it is natural to investigate and apply the lessons of primitive-based stereo matching in a multimodal context. Seminal works in multiprimitive stereo, such as the approach developed by Marapane [24], will serve as guide for our investigation and development of a framework for a multimodal stereo algorithm. We will describe an approach to region-based matching in Section III.C.1.

In order to analyze the multimodal imagery and offer a direct comparison to both unimodal color and unimodal infrared stereo setups, we have designed a testbed capable of generating the three separate, yet synchronized, stereo imagery. Utilizing a two color, two infrared system and a four-input frame grabber, we are able to obtain synchronized uncompressed streams from each camera. The cameras have been arranged and aligned carefully on a metal frame that supports variable baselines and

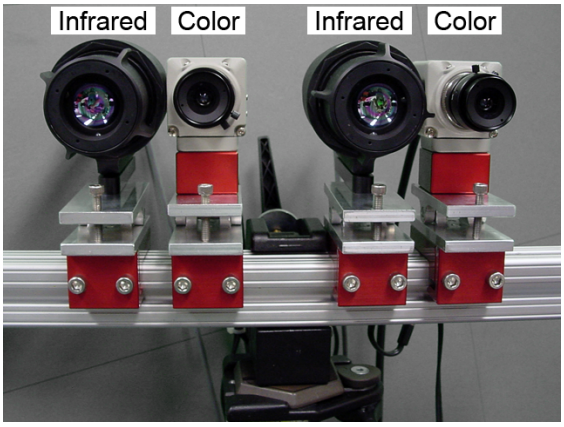
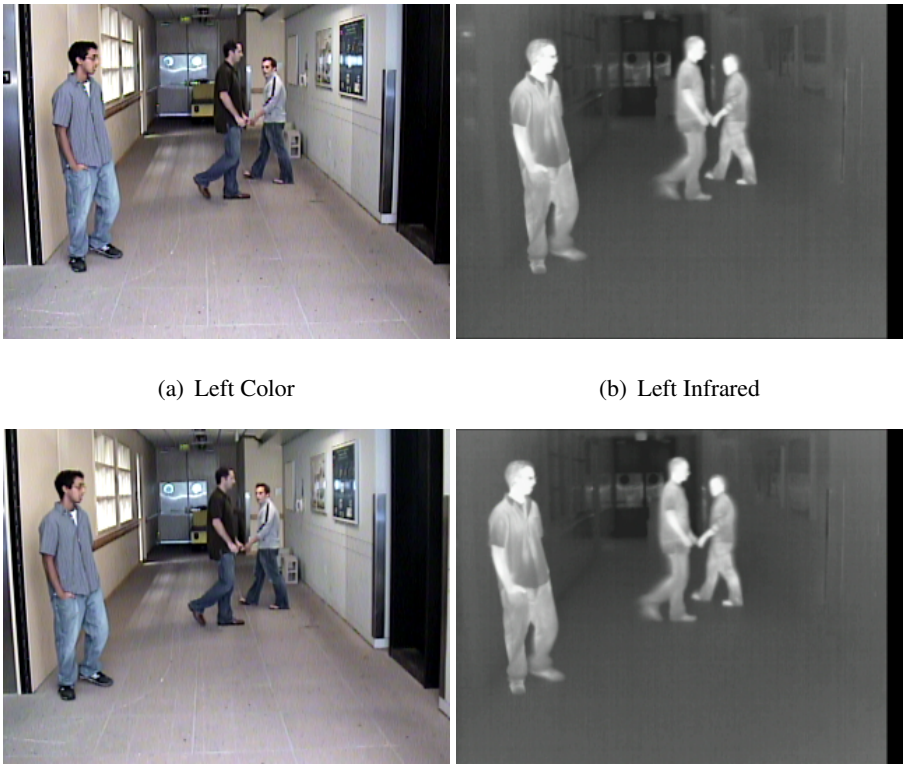


Figure III.3: Experimental Testbed

easy addition, removal and adjustment of each camera (Figure III.3). The cameras can be calibrated using a single calibration board to yield rectification parameters for color, thermal and multimodal stereo pairs. Once calibrated, it is quite simple and quick to conduct experiments in a manner that can yield frame-by-frame comparison of results across individual stereo rigs.

III.C.1 Region-based Cross-Spectral Stereo Matching using Disparity Voting

Resolving stereo correspondences through regions is one of the classical approaches to utilizing image features for image matching. Traditionally, works such as those by Marapane [25] and Cohen *et al.* [26] use image segmentation to obtain regions and can achieve a coarse disparity estimate. Usually this sort of approach is one part of a larger stereo matching algorithm with the coarse disparity map used to guide refinements at finer detail. More recently, approaches that use the concept of over-segmentation have been applied to stereo imagery [27], [28]. By over-segmenting the image into very small regions, matching can be done in a progressive manner similar to pixel-based energy minimization functions. These over-segmentation approaches rely on the intensity similarity properties of unimodal stereo imagery and are therefore not readily extendable to the multimodal case. The challenge in applying region-based approaches to multimodal imagery lies in finding region segmentation that yields small



(a) Left Color (b) Left Infrared

(c) Right Color (d) Right Infrared

Figure III.4: Example raw captured imagery from testbed.

enough regions to allow for a fine level of disparities while maintaining large enough regions to allow for reliable and robust matching.

Our registration algorithm [20] addresses the registration of objects at different depths in relatively close range surveillance scenes. It eliminates the need for perfectly segmented bounding boxes by relying on reasonable initial foreground segmentation and using a disparity voting algorithm to resolve the registration for occluded or malformed segmentation regions. This approach gives robust registration disparity estimation with statistical confidence values for each estimate. Figure VI.3 shows a flowchart outlining our algorithmic framework. Individual modules are described in the subsequent sections.

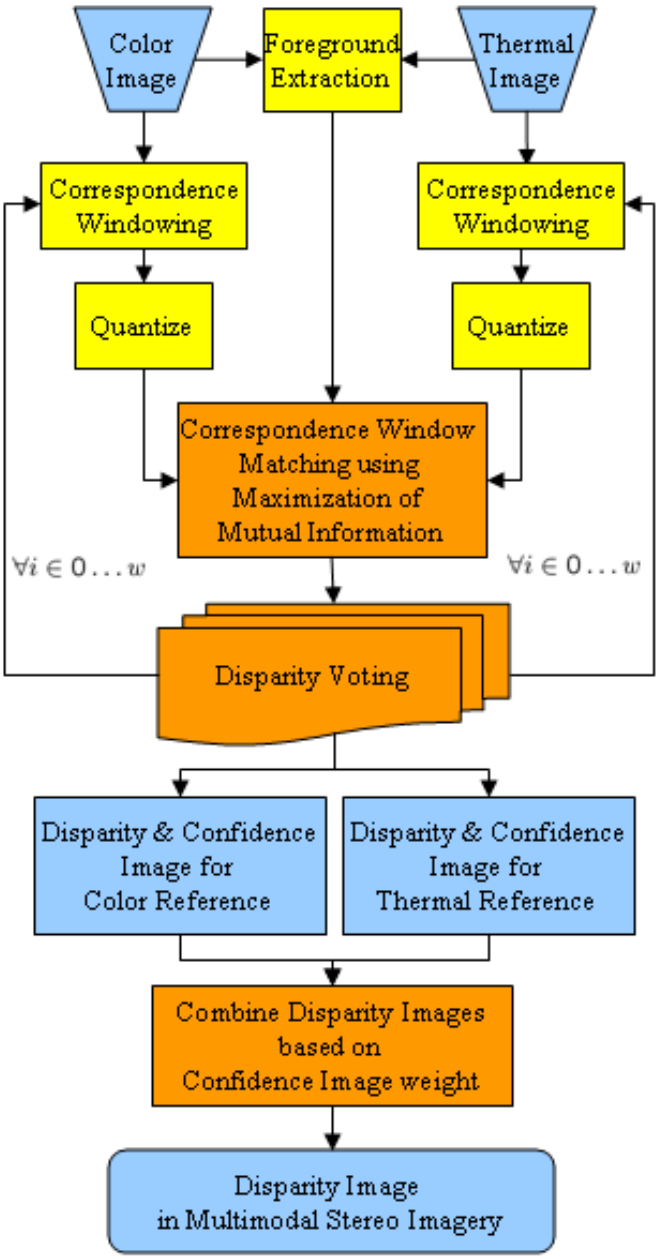


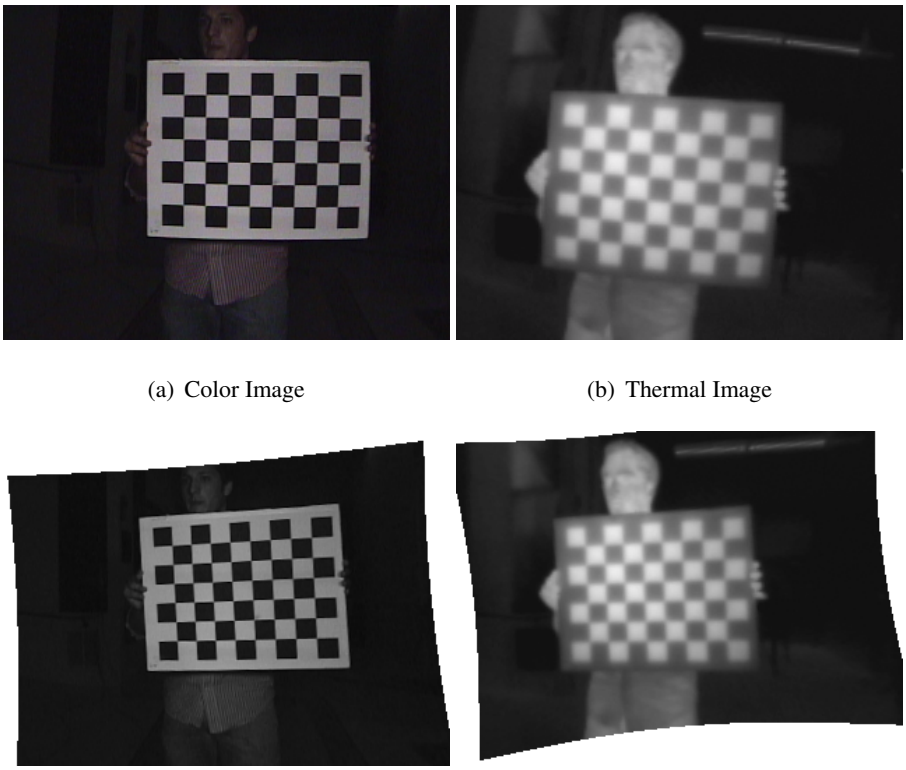
Figure III.5: Flowchart of disparity voting approach to multimodal image registration.

Multimodal Image Calibration

A minimum camera solution for registering multimodal imagery in these short range surveillance situations would be to use a single camera from each modality, arranged in a stereo pair. Unlike colocating the cameras, arranging the cameras into a stereo pair allows objects at different depths to be registered. To perform this type of registration, it is desirable to first calibrate the color and thermal infrared cameras. Knowing the intrinsic and extrinsic calibration parameters transforms the epipolar lines to lie along the image scanlines, enabling disparity correspondence matching to be a one-dimensional search. Calibration can be performed using standard techniques, such as those available in the Camera Calibration Toolbox for Matlab [29]. The toolbox assumes input images from each modality where a calibration board is visible in the scene. In typical visual setups, this is simply a matter of placing a checkerboard pattern in front of the camera. However, due to the large differences in visual and thermal imagery, some extra care needs to be taken to ensure the calibration board looks similar in each modality. A solution is to use a standard calibration board and illuminate the scene with high intensity halogen bulbs placed behind the cameras. This effectively warms the checkerboard pattern, making the visually dark checks appear brighter in the thermal imagery. Placing the board under constant illumination reduces the blurring associated with thermal diffusion and keeps the checkerboard edges sharp, allowing for calibration with subpixel accuracy. An example pair of images in the visual and thermal infrared domain and the subsequently calibrated and rectified image pair is shown in Figure III.6.

Image Acquisition and Foreground Extraction

The acquired and rectified image pairs are denoted as I_L, the left color image, and I_R, the right thermal image. Due to the high differences in imaging characteristics, it is very difficult find to correspondences for the entire scene. Instead, registration is focused on the pixels that correspond to foreground objects of interest. Naturally then, it is desirable to determine which pixels in the frame belong to the foreground. In this step, only a rough estimate of the foreground pixels is necessary and a fair amount of



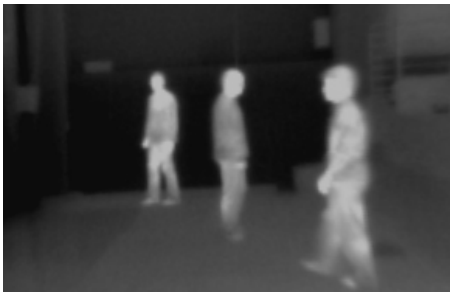
(a) Color Image (b) Thermal Image

(c) Rectified Color Image (d) Rectified Thermal Image

Figure III.6: Multimodal Stereo Calibration using a heated calibration board to allow for a visible checkerboard pattern in thermal imagery

(a) Color

(b) Color Segmentation



(c) Thermal

(d) Thermal Segmentation

Figure III.7: Image acquisition and foreground extraction for color and thermal imagery.

false positives and negatives is acceptable. Any "good" segmentation algorithm could potentially be used with success. The corresponding foreground images are F_L and F_R, respectively. Additionally, the color image is converted to grayscale for mutual information based matching. Example input images and foreground maps are shown in Figure III.7.

Correspondence Matching using Maximization of Mutual Information

Once the foreground regions are obtained, the correspondence matching can begin. Matching occurs by fixing a correspondence window along one reference image in the pair and sliding the window along the second image to find the best match. Let h and w be the height and width of the image, respectively. For each column $i \in 0 \ldots w$, let $W_{L,i}$ be a correspondence window in the left image of height h and width M centered on column i. The width M that produces the best results can be experimentally determined for a given scene. Typically, the value for M is significantly less than the

width of an object in the scene. Define a correspondence window $W_{R,i,d}$ in the right image having height h^*, the largest spanning foreground distance in the correspondence window, and centered at a column $i + d$, where d is a disparity offset. For each column i, a correspondence value is found for all $d \in d_{\min} \ldots d_{\max}$.

Given the two correspondence windows $W_{L,i}$ and $W_{R,i,d}$, we first linearly quantize the image to N levels such that

$$N \approx \sqrt{Mh^*/8} \tag{III.9}$$

where Mh^* is the area of the correspondence window. The result in (III.9) comes from Thevenaz and Unser's [6] suggestion that this equation is reasonable to determine the number of levels needed to give good results for maximizing the mutual information between image regions.

Now we can compute the quality of the match between the two correspondence windows by measuring the mutual information between them. The mutual information between two image patches is defined as

$$I(L, R) = \sum_{l,r} P_{L,R}(l, r) \log \frac{P_{L,R}(l, r)}{P_L(l) P_R(r)} \tag{III.10}$$

where $P_{L,R}(l, r)$ is the joint probability mass function (pmf) and $P_L(l)$ and $P_R(r)$ are the marginal pmf's of the left and right image patches, respectively.

The two-dimensional histogram, g, of the correspondence window is utilized to evaluate the pmf's needed to determine the mutual information. The histogram g is an N by N matrix so that for each point, the quantized intensity levels l and r from the left and right correspondence windows increment $g(l, r)$ by one. Normalizing by the total sum of the histogram gives the probability mass function

$$P_{L,R}(l, r) = \frac{g(l, r)}{\sum_{l,r} g(l, r)} \tag{III.11}$$

The marginal probabilities can be easily determined by summing $P_{L,R}(l, r)$ over the appropriate dimension.

$$P_L(l) = \sum_r P_{L,R}(l,r) \quad \text{(III.12)}$$

$$P_R(r) = \sum_l P_{L,R}(l,r) \quad \text{(III.13)}$$

Now that we are able to determine the mutual information for two generic image patches, let's define the mutual information between two specific image patches as $I_{i,d}$ where again i is the center of the reference correspondence window and $i+d$ is the center of the second correspondence window. For each column i, we have a mutual information value $I_{i,d}$ for $d \in d_{\min} \ldots d_{\max}$. The disparity d_i^* that best matches the two windows is the one that maximizes the mutual information

$$d_i^* = \arg\max_d I_{i,d} \quad \text{(III.14)}$$

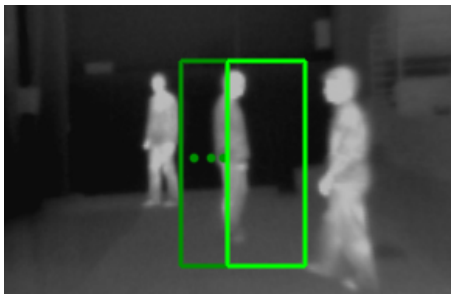
The process of computing the mutual information for a specific correspondence window is illustrated in Figure III.8. An example plot of the mutual information values over the range of disparities is also shown. The red box in the color image is a visualization of a potential reference correspondence window. Candidate sliding correspondence windows for the thermal image are visualized in green boxes.

Disparity Voting with Sliding Correspondence Windows

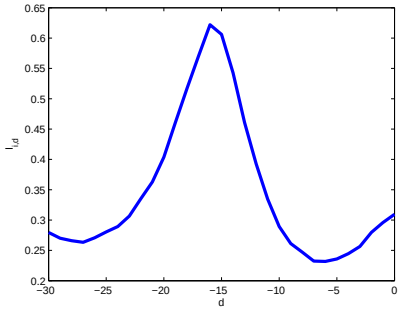
We wish to assign a vote for d_i^*, the disparity that maximizes the mutual information, to all foreground pixels in the reference correspondence window. Define a disparity voting matrix D_L of size $(h, w, d_{\max} - d_{\min} + 1)$, the range of disparities. Then given a column i, for each image pixel that is in the correspondence window and foreground map, $(u,v) \in (W_{L,i} \cap F_L)$, we add to the disparity voting matrix at $D_L(u,v,d_i^*)$.

Since the correspondence windows are M pixels wide, pixels in each column in the image will have M votes for a correspondence matching disparity value. For each pixel (u,v) in the image, D_L can be thought of as a distribution of matching disparities from the sliding correspondence windows. Since it is assumed that all the pixels

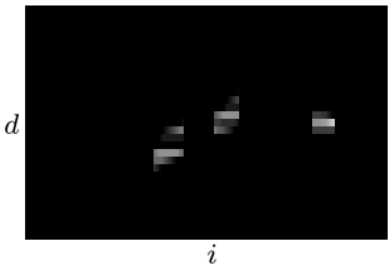
(a) Color Image



(b) Thermal Image



(c) Mutual Information



(d) Disparity Voting Matrix

Figure III.8: Mutual Information for Correspondence Windows.

attributed to a single person are at the same distance from the camera, a good match should have a large number of votes for a single disparity value. A poor match would be widely distributed across a number of different disparity values. Figure III.8(d) shows the disparity voting matrix for a sample row in the color image. The x-axis of the image is the column number i of the input image. The y-axis of the image is the range of disparities $d = d_{\min} \ldots d_{\max}$, which can be experimentally determined based on scene structure and the areas in the scene where activity will occur. Entries in the matrix correspond to the number of votes given to a specific disparity at a specific column in the image. Brighter areas correspond to a higher vote tally.

The complementary process of correspondence window matching is also performed by keeping the right thermal infrared image fixed. The algorithm is identical to the one described above, switching the left and right denotations. The corresponding disparity accumulation matrix is given as D_R.

Once the disparity voting matrices have been evaluated for the entire image, the final disparity registration values can be determined. For both the left and right images, we determine the best disparity value and its corresponding confidence measure as

$$D_L^*(u,v) = \arg\max_d D_L(u,v,d) \tag{III.15}$$

$$C_L^*(u,v) = \max_d D_L(u,v,d) \tag{III.16}$$

For a pixel (u,v) the values of $C_L^*(u,v)$ represent the number of times the best disparity value $D_L^*(u,v)$ was voted for. A higher confidence value indicates that the disparity maximized the mutual information for a large number of correspondence windows and in turn, the disparity value is more likely to be accurate than at a pixel with lower confidence. Values for D_R^* and C_R^* are similarly determined. The values of D_R^* and C_R^* are also shifted by their disparities so that they align to the left image:

$$D_S^*(u, v + D_R^*(u, v)) = D_R^*(u, v) \tag{III.17}$$

$$C_S^*(u, v + D_R^*(u, v)) = C_R^*(u, v) \tag{III.18}$$

Once the two disparity images are aligned, they can be combined. We experimented with various combination approaches, including boolean OR and AND operations. Our experiments indicated that using an AND operation yielded the best overall registration on our test examples. So for all pixels (u, v) such that $C_L^*(u, v) > 0$ and $C_S^*(u, v) > 0$,

$$D^*(u, v) = \begin{cases} D_L^*(u, v), & C_L^*(u, v) \geq C_S^*(u, v) \\ D_S^*(u, v), & C_L^*(u, v) < C_S^*(u, v) \end{cases} \tag{III.19}$$

The resulting image $D^*(u, v)$ is the disparity image for all the overlapping foreground object pixels in the image. It can be used to register multiple objects in the image, even at very different depths from the camera. Figure III.9 shows the result of registration for the example frame carried throughout the algorithmic derivation. Figure III.9(a) shows the computed disparity image D^*, while Figure III.9(b) shows the initial alignment of the color and thermal images and Figure III.9(b) shows the alignment after shifting the foreground pixels by the resulting disparity image. The thermal foreground pixels are overlaid (in green) on the color foreground pixels (in purple).

The resulting registration in Figure III.9 is successful in aligning the foreground areas associated with each of the three people in the scene. Each person in the scene lies at a different distance from the camera and yields a different disparity value that will align its corresponding image components.

III.D Summary

In this chapter we have detailed the issues and challenges in finding measures of similarity for solving stereo correspondences in multimodal imagery. While

(a) Disparity Image (b) Unregistered (c) Registered

Figure III.9: The resulting disparity image D^* from combining the left and right disparity images D_L^* and D_S^* as defined in (III.19).

recent energy minimization approaches have used mutual information to solve pointwise stereo correspondences for a wide variety of synthetically altered imagery, our experiments demonstrate that these approaches are unable to resolve the stereo correspondences for the true multimodal nature of color and thermal images. We have shown that because the intensities associated with color and thermal imagery are uncorrelated, using mutual information as a similarity measure will not yield good disparity estimates in an energy minimization framework.

We developed a method for analyzing region-based similarity and performed extensive experiments that demonstrate the ability to provide robust disparity estimates in multimodal imagery. The disparity voting algorithm we present can successfully register multiple objects in the scene that lie at different depths from the camera. Such scenes are common to person-centric vision applications such as surveillance [30] and pedestrian detection [31], which we will explore in the next chapter. By describing a general framework and providing a discussion of the issues and challenges inherent in developing such a system, we have introduced a successful algorithm and laid the groundwork for future research and advancement in multimodal stereo correspondence matching.

The text of this chapter, in part, is a reprint of the material as it appears in: Stephen J. Krotosky and Mohan M. Trivedi, “Mutual Information Based Registration of Multimodal Stereo Videos for Person Tracking” in *Computer Vision and Image Un-*

derstanding, Special Issue on Advances in Vision Algorithms and Systems Beyond the Visible Spectrum, Vol. 106, Issues 2-3, May-June 2007 and Stephen J. Krotosky and Mohan M. Trivedi, "Registration of Multimodal Imagery with Occluding Objects using Mutual Information", *Applied Perception in Thermal Infrared Imagery*, in press. I was the primary researcher of the cited material and the co-author listed in these publications directed and supervised the research which forms a basis for this chapter.

Chapter IV

Evaluation of Cross-Spectral Stereo Registration

IV.A Introduction

We evaluate the disparity voting registration algorithm using color and thermal data for a variety of application scenarios. Oriented in the same direction with a baseline of 10 cm, the cameras were placed so that the optical axis was approximately parallel to the ground. This placement was used to satisfy the assumption that there would be approximately constant disparity across all pixels associated with a specific person in the frame. Placing the cameras in this sort of position is a reasonable thing to do, and such a position is appropriate for many applications including surveillance and pedestrian detection.

IV.B Indoor Surveillance Experiments

Video was captured as up to four people moved throughout an indoor environment designed to mimic an indoor surveillance scenario. For these specific experiments, foreground segmentation in the visual imagery was done using the codebook model proposed by Kim, *et al.* [32]. In the thermal imagery, the foreground is obtained using an intensity threshold under the assumption that the people in the foreground are hotter than the background. This approach provided reasonable segmentation in each image. In cases where segmentation can only be obtained for one modality, the disparities can be computed with only that modality as the reference, at the cost of less robustness. We will show successful registration for examples of varying segmentation quality. The goal was to obtain registration results for various configurations of people including different positions, distances from camera, and levels of occlusion.

Examples of successful registration are shown in Figure IV.1. Columns (a) and (b) show the input color and thermal images, while column (c) illustrates the initial registration of the objects in the scene and column (d) shows the resulting registration overlay after the disparity voting has been performed. These examples show the registration success of the disparity voting algorithm in handling occlusion and properly registering multiple objects at widely disparate depths from the camera.

(a) Color (b) Infrared (c) Unregistered (d) Registered

Figure IV.1: Registration results using Disparity Voting Algorithm for example frames.

(a) Well Aligned (b) Partially Aligned (c) Misaligned

Figure IV.2: Examples of good and bad registration alignment in our evaluation. Bad alignments are highlighted in red.

Table IV.1: Registration Results for Disparity Voting Algorithm with Multiple People in a Scene

No. Objects in Frame	No. Frames Correct	Total Frames	% Correct
1	55	55	100.00 %
2	171	172	99.42 %
3	1087	1111	97.84 %
4	690	720	95.83 %
Total	2003	2058	97.33 %

We have analyzed the registration results of our disparity voting algorithm for more than 2000 frames of captured video. To evaluate the registration, we define a correct frame as when the color and infrared data corresponding to each foreground object in the scene were visibly aligned. If one or more objects in the scene are not visibly aligned, then the registration is deemed incorrect for the entire frame. We evaluate an object as aligned by examining the alignment of body parts such as the head and limbs where the amount of overlay can be visibly seen. We enforce that all parts of the person be aligned and any missing or misalignment is deemed as unaligned. Figure IV.2 shows examples of good and bad alignments in our evaluation. Table VI.1 shows the results of this evaluation. The data is broken down into groups based on the number of objects in the scene.

This analysis shows that when there was no visible occlusion in the scene, registration was correct 100% of the time. We further break down the analysis to consider

Table IV.2: Registration Results for Disparity Voting Algorithm with Multiple People in a Scene: Frames with Occlusion

No. Objects in Frame	No Frames Correct	Total Frames	% Correct
2	51	52	98.08 %
3	653	677	96.45 %
4	581	611	95.09 %
Total	1285	1340	95.90 %

only the frames where there are occluding objects in the scene. Under these conditions, the registration success of the disparity voting algorithm is shown in Table VI.2. The registration results for the occluded frames are still quite high, with most errors occurring during times of near total occlusion.

IV.C Outdoor Pedestrian Detection Experiments

To obtain segmentation in a moving vehicle for pedestrian detection, we need to modify the algorithm slightly to achieve the initial segmentation. We use an optical flow-based approach to detect moving pedestrians in the scene [33]. Our experiments have shown this approach is relatively robust at low speeds (< 10mph) and could be adapted for higher speeds with egomotion estimation. Low speed analysis is useful in a variety of driving scenarios, including parking lots, residential and shopping areas, and starting or stopping at a traffic signal. Additionally, while stationary pedestrians pose a segmentation issue for optical flow techniques, we expect static objects can be identified through long term tracking of the scene.

Given the optical flow estimates for motion in the horizontal m_u and vertical m_v directions as well as occluded regions m_{occ}, we estimate foreground regions F where there is motion in either the horizontal or vertical direction and no occlusion as shown in (IV.1). Morphological operations smooth the estimate.

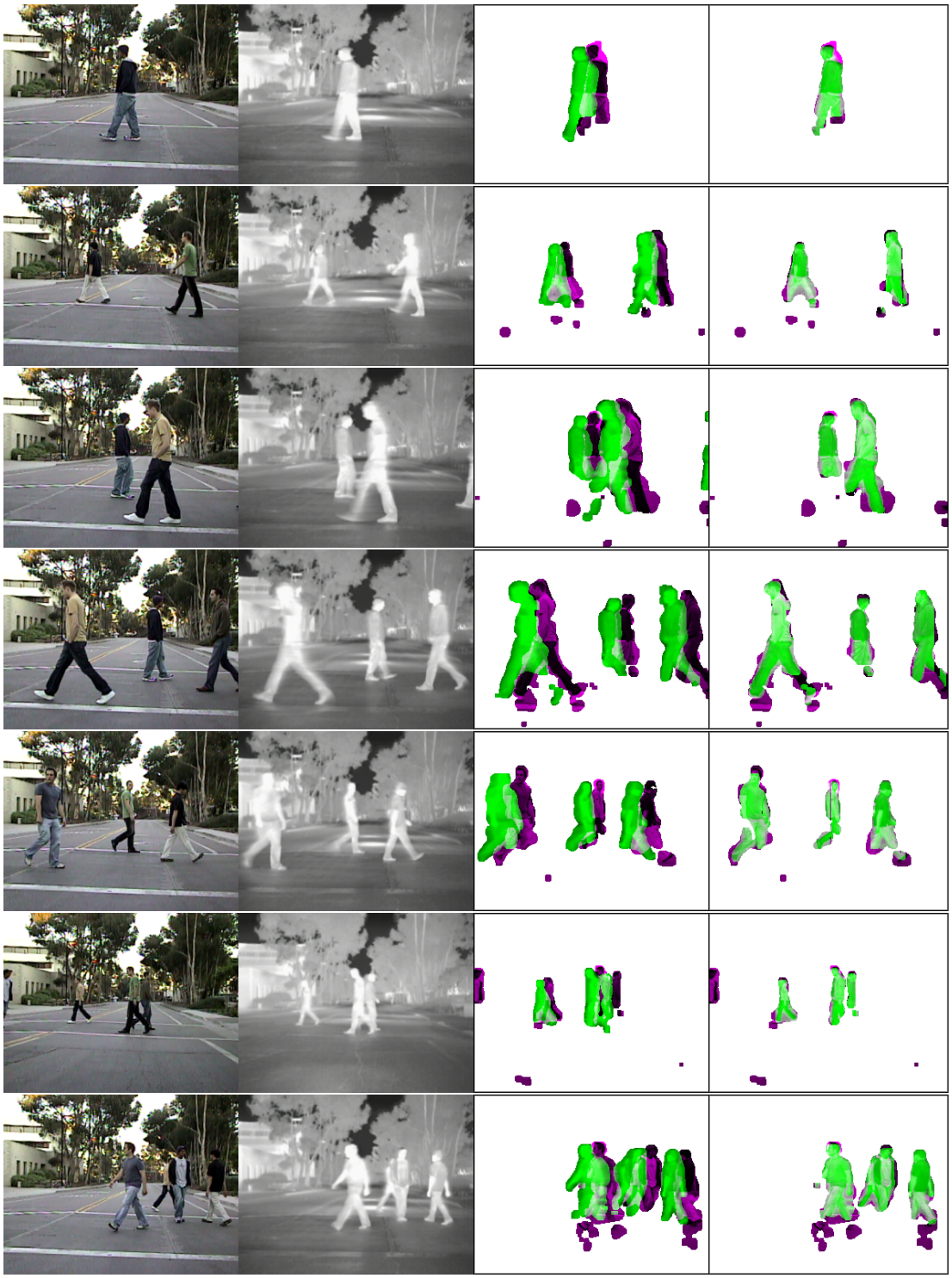
$$F = \Big((|m_u| > 0) \cup (|m_v| > 0) \Big) \cap (m_{occ} = 0) \tag{IV.1}$$

Table IV.3: Cross-Spectral Stereo Registration of Pedestrian Regions

# Peds	Peds Correct	Total	% Correct
1	126	145	86.9%
2	805	942	85.5%
3	633	744	85.1%
4	96	108	88.9%
Total	1690	1939	87.2%

We analyzed the ability of the cross-spectral stereo correspondence matching algorithm to match pedestrian regions in an outdoor experimental environment. Experiments were conducted at mid-afternoon on a sunny day from a camera mounted to our moving LISA-P test vehicle. The goal was to obtain successful correspondence matching for various configurations of people including different positions, distances from the camera and levels of occlusion. We evaluate the success of the correspondence algorithm by visually inspecting the alignment of the corresponding color and infrared pedestrian regions. If the regions are visually well-aligned, then the correspondence is considered correct. If the regions are misaligned, missing or only partially aligned, the correspondence is deemed incorrect. Table VI.1 summarizes the results for our experiments and Figure IV.3 shows examples of correct correspondence matching. Additional experiments [31] demonstrate the robustness of the approach to different capture devices and environmental conditions.

One challenge associated with this approach to cross-spectral stereo lies in the vertical artifacts generated from the multiple voting windows. The resulting registration disparities often have hard vertical edges in disparity discontinuities. This is especially evident when there are occluding pedestrians in the scene, as there is an inherent disparity discontinuity that is forced to be a vertical edge. Figure IV.4 illustrates an example of this artifacting error. Despite these errors, we are still able to identify two distinct obstacle regions. Additionally, it is possible that even good registration results can be off by a pixel in either direction because of the integer-only disparity matching of our approach. By incorporating an approach with subpixel accuracy, the registration and its



(a) Color (b) Infrared (c) Unaligned (d) Aligned

Figure IV.3: Cross-Spectral Stereo Registration Results for Pedestrian Detection

(a) Color (b) Infrared (c) Disparity (d) Aligned

Figure IV.4: Disparity discontinuity errors in cross-spectral stereo analysis due to artifacting arising from windowed correspondence matching.

ability to produce depth estimates will be more robust.

The requirement of an initial segmentation, while necessary for success in this algorithm, is limiting in several aspects. First, segmentation is a challenging task and the result can often be noisy or can easily over or under estimate the true object boundaries, leading to registration errors. The motivation behind the initial segmentation is to provide some regions appropriately sized for matching features in the color and infrared imagery. However, the very idea of an initial segmentation precludes registration estimates for regions not within the segmentation boundaries. Clearly a better approach would be to register the features in the whole image without the segmentation requirement. Achieving this is an open research challenge that we are actively pursuing. We feel that a multi-feature matching approach that can integrate structural feature matching, such as edges, with pixel or area based matching may yield improved results.

IV.D Accuracy Evaluation using Ground Truth Disparity Values

In order to demonstrate the accuracy of our disparity voting algorithm (DV) in handling occlusions, we offer a quantitative comparison to ground truth. It is our contention that the disparity voting algorithm will provide good registration results during occlusions, when initial segmentation gives regions that contained merged objects. Our disparity voting algorithm makes no assumptions about the assignment of pixels to individual objects, only that a reasonable segmentation can be obtained. We demonstrate

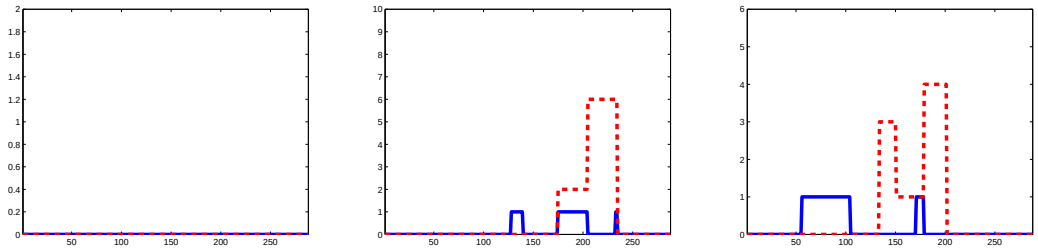
(a) Ground Truth (b) BB Disparity (c) DV Disparity

Figure IV.5: Comparison of Bounding Box (BB) approach to the proposed Disparity Voting algorithm for ground truth segmentation.

that the disparity voting registration can successfully register all objects in the scene even through occlusions. We will also show the results for bounding box approaches (BB) [17] for completeness.

We generate the ground truth by manually segmenting the regions that correspond to foreground for each image. We then determine the ground truth disparity by individually matching each manually segmented object in the scene. This ground truth disparity image allows us to directly and quantitatively compare the registration success of the disparity voting algorithm and the bounding box approach. By comparing the registration results to the ground truth disparities, we are able to quantify the success of each algorithm and show that the disparity voting algorithm outperforms the bounding box approach for occluding object regions.

Figure IV.5 illustrates the ground truth disparity comparison tests. Column (a) shows the ground truth disparity, column (b) shows the disparity generated using the



(a) Figure IV.5.1 (b) Figure IV.5.2 (c) Figure IV.5.3

Figure IV.6: Plots of $|\Delta D|$ from ground truth for each example in Figure IV.5. Bounding Box errors for an example row are plotted in dotted red, while errors in Disparity Voting registration are plotted in solid blue.

bounding box (BB) algorithm, and column (c) shows the disparity generated using the disparity voting (DV) algorithm. Figure IV.6 plots the absolute difference in disparity values ($|\Delta$ Disparity$|$) from the ground truth for each corresponding row in Figure IV.5. The BB results are plotted in dotted red, while the DV results are plotted in solid blue. Notice how the two algorithms perform identically to ground truth in the first row, as there are no occlusion regions. The subsequent examples all have occlusion regions and the DV approach more closely follows ground truth than the BB approach. The BB registration results have multiple objects registered at the same depth though the ground truth shows that they are at separate depths. Our disparity voting algorithm is able to determine the distinct ground truth disparities for different objects and the $|\Delta$ Disparity$|$ plots show that the DV algorithm is quantitatively closer to the ground truth, and with most registration errors within one pixel of ground truth with larger errors usually occurring only in small portions of the image. On the other hand, when errors occur in the bounding box approach, the resulting disparity offset error is large and occurs for the entire scope of the erroneously registered object.

IV.E Comparative Study of Registration Algorithms with Non-Ideal Segmentation

We perform a qualitative evaluation using the real segmentations generated from codebook background subtraction in the color image and intensity thresholding in the thermal image. These common segmentation algorithms only give foreground pixels and make no attempt to discern the structure of objects in the scene. Figure IV.7 illustrates several examples that compare the registration results of the disparity voting and bounding box algorithms. Notice how the disparities for the bounding box (BB) algorithm in row (5) are constant for the entire occlusion region even though the objects are clearly at very different disparities. The disparity results for our disparity voting algorithm in row (6) show distinct disparities in the occlusion regions that correspond to the appropriate objects in the scene. Visual inspection of rows (7) and (8) show that the resulting registered alignment from the disparity values is more accurate for the DV approach.

Figure IV.8 shows the registration alignment for each algorithm in closer detail for a selection of frames. Notice how the disparity voting approach is able to align each object in the frame, while the bounding box approach has alignment errors due to the fact that the segmentation of the image yielded bounding boxes that contained more than one object. Clearly, disparity voting is able to handle the registration in these occlusion situations and the resulting alignment appears qualitatively better than the bounding box approach.

IV.F Robustness Evaluation

We demonstrate the robustness of our algorithm by applying it to another set of data taken of a different scene with a different set of cameras. For these experiments, we have up to 6 people move through an approximately 6m x 6m environment. The cameras are arranged with a 10 cm baseline and are calibrated and rectified as described in Section III.C.1. Again, segmentation is performed using the codebook background

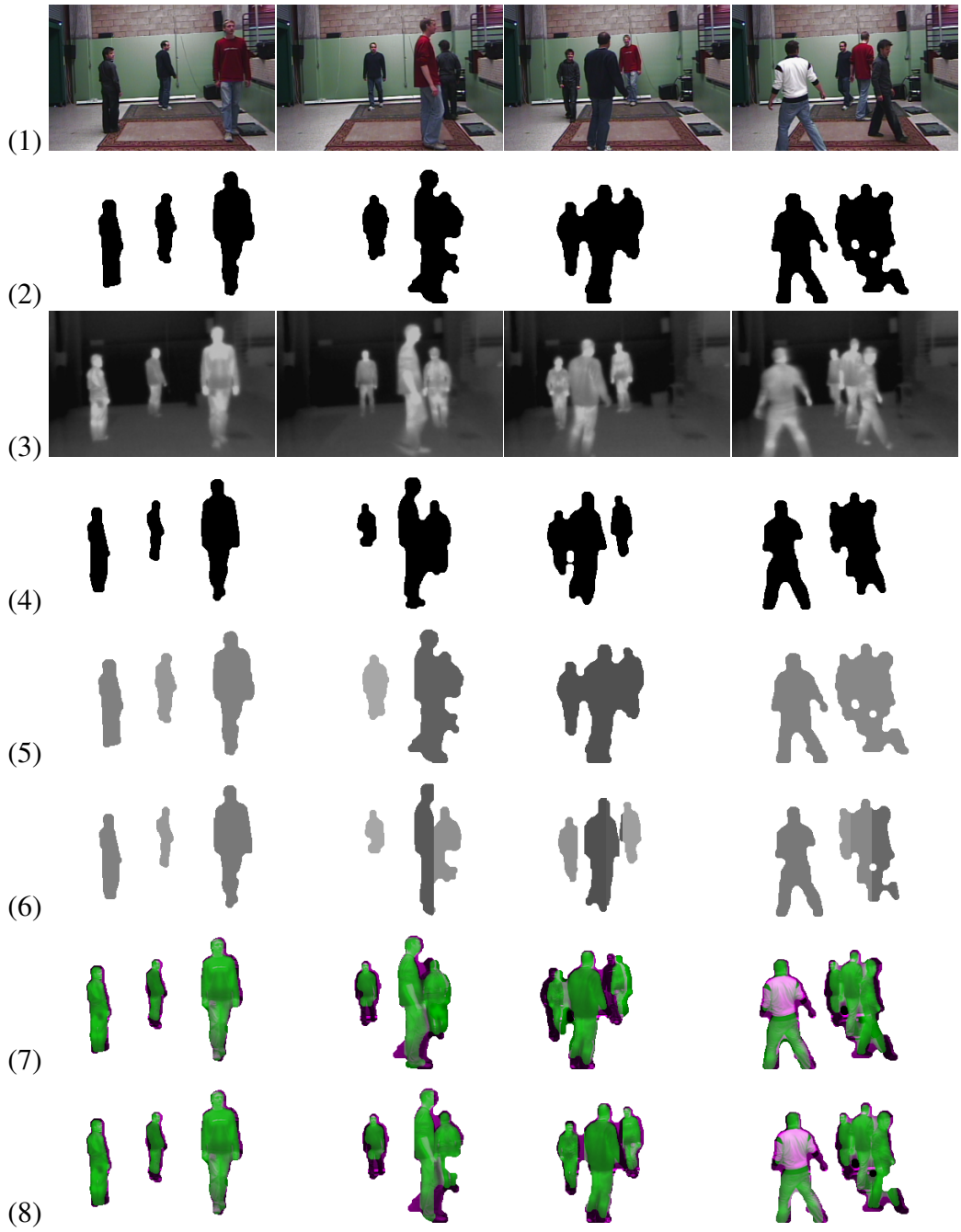
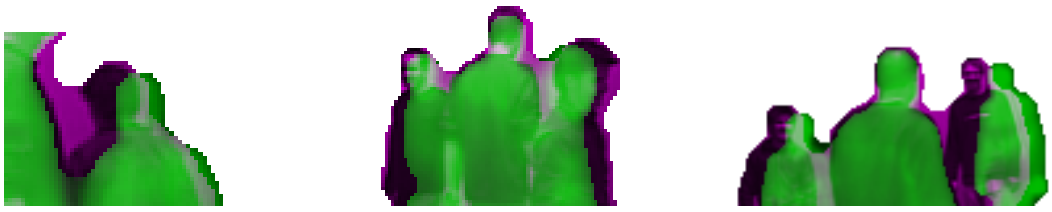


Figure IV.7: Comparison of BB algorithm to the proposed Disparity Voting (DV) algorithm for a variety of occlusion examples using non-ideal segmentation: (1) the color image, (2) the color segmentation, (3) the thermal image, (4) the thermal segmentation, (5) the BB Disparity Image, (6) the DV Disparity Image, (7) the BB Registration, (8) the DV Registration.



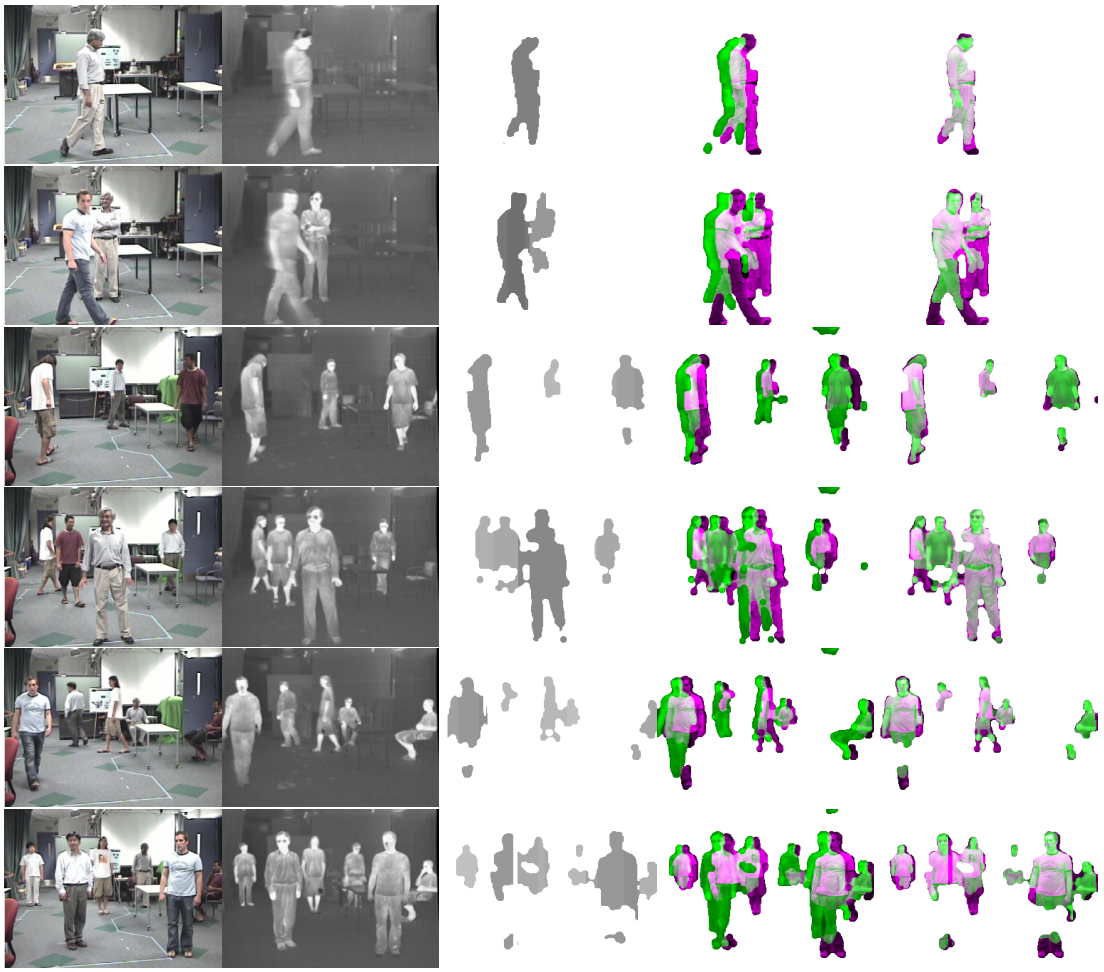
(a) BB Registration

(b) DV Registration

Figure IV.8: Details of registration alignment errors in the bounding box registration approach and corresponding alignment success for the Disparity Voting (DV) Algorithm for several occlusion examples using non-ideal segmentation.

model for the color imagery and intensity thresholding for the thermal imagery. Correspondence window sizes and threshold values were kept constant from past experiments.

Figure IV.9 shows successful registration for example frames containing an increasing number of people in the scene. Column (c) of the figure shows distinct levels of alignment disparity for each person in the scene and column (e) shows the resulting registered alignment. Notice how the disparity voting algorithm is able to properly determine the disparities necessary to align the color and thermal image in situations with multiple people and multiple levels of occlusion. Figure IV.10 shows detailed examples of the registration alignment. Note how image features, especially facial region, appear well aligned in the images.



(a) Color (b) Thermal (c) Disparity (d) Unregistered (e) Registered

Figure IV.9: Examples illustrating the robustness of the disparity voting algorithm in registering multiple people in a scene. Each row contains an increasing number of people. Column (e) illustrates the registration using disparity voting. It is a marked improvement over the initial, unregistered image in column (d).

Figure IV.10: Detailed examples of successful registration alignment using disparity voting.

IV.G Multimodal Video Analysis for Person Tracking: Basic Framework and Experimental Study

We have shown that the disparity voting algorithm for multimodal registration is a robust approach to estimating the alignment disparities in scenes with multiple occluding people. The disparities generated from the registration process yield values that can be used to differentiate the people in the room. It is with this in mind that we investigate the use of multimodal disparity as a feature for tracking people in a scene.

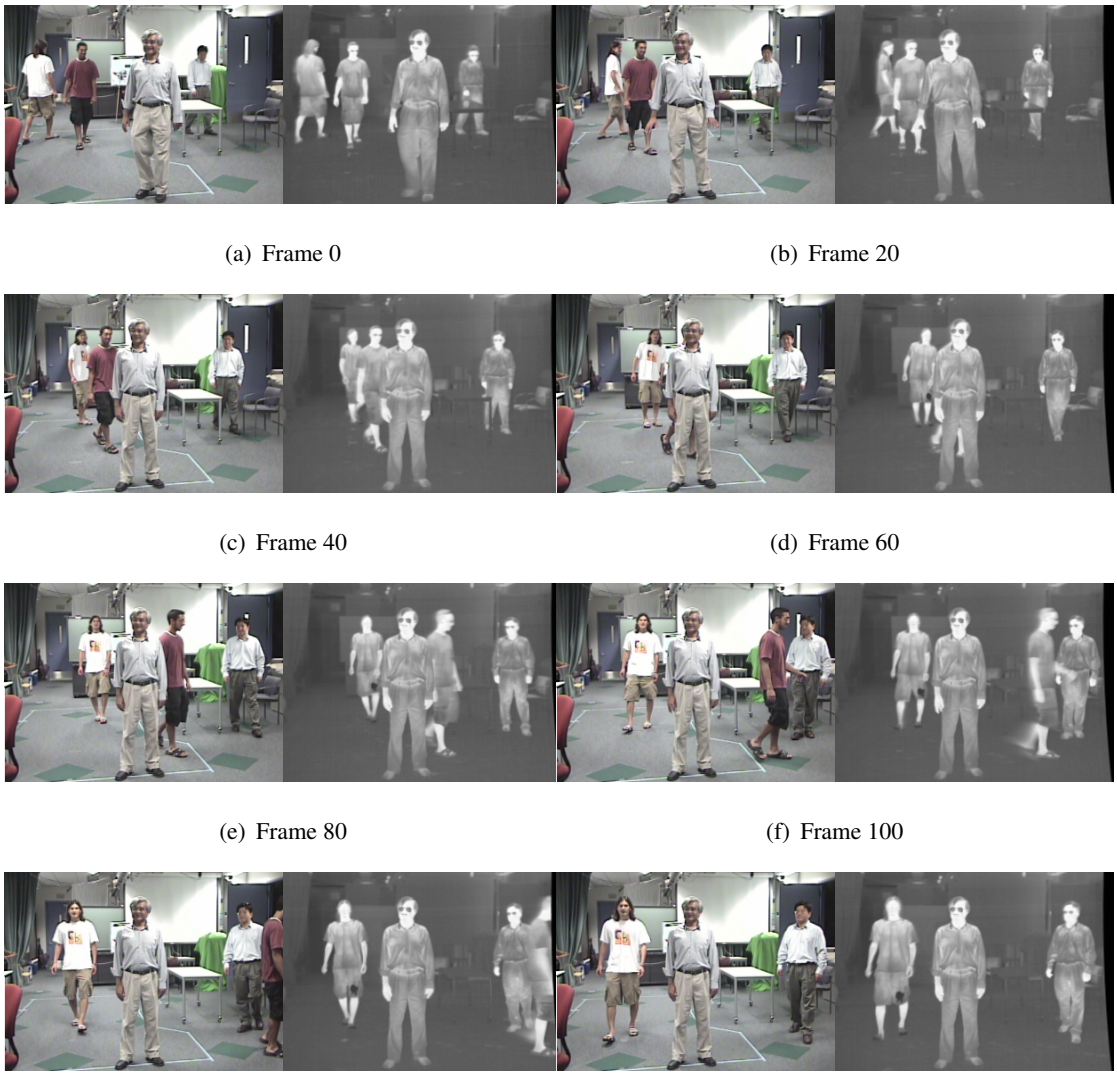
Tracking human motion using computer vision approaches is a well-studied area of research and a good survey by Moeslund and Granum [34] gives lucid insight into the issues, assumptions and limitations of a large variety of tracking approaches. One approach, disparity based tracking, has been investigated for conventional color stereo cameras and has proven quite robust in localizing and maintaining tracks through occlusion, as the tracking is performed in 3D space by transforming the stereo image estimates into a plan-view occupancy map of the imaged space [35]. We wish to explore the feasibility of using such approaches to tracking with the disparities generated from disparity voting registration. An example sequence of frames in Figure IV.11 illustrates the type of people movements we aim to track. The sequence has multiple people occupying the imaged scene. Over the sequence, the people move in a way where there are multiple occlusions of people at different depths. The registration disparities that are

used to align the color and thermal images can be used as a feature for tracking people through these occlusions and maneuvers.

Figure IV.12 shows an algorithmic framework for multimodal person tracking. In tracking approaches, representative features are typically extracted from all available images in the setup [36]. Features are used to associate tracks from frame to frame and the output of the tracker is often used to guide subsequent feature extraction. All of these algorithmic modules are imperative for reliable and robust tracking. For our initial investigations, we will focus on the viability of registration disparity as a tracking feature.

In order to determine the accuracy of the disparity estimates for tracking, we first calibrate the scene. This is done by having a person walk around the testbed area, stopping at preset locations in the scene. At each location we measure the disparity generated from our algorithm and use that as ground truth for analyzing the disparities generated when there are more complex scenes with multiple people and occlusions. Figure IV.13(a) is the variable baseline multimodal stereo rig and Figure IV.13(b) shows the ground truth disparity range for the testbed from the calibration experiments captured with this rig.

To show the viability of registration disparity as a tracking feature in a multimodal stereo context, we compare ground truth positional estimates to those generated from the disparity voting algorithm. Lateral position information for each track was hand segmented by clicking on the center point of the person's head in each image. This is a reasonable method, as robust head detection algorithms for head detection could be implemented for both color and thermal imagery (skin-tone, hot spots, head template matching). Approaches such as vertical projection or v-disparity could also be used to determine the locations of people in the scene. Ground truth disparity estimates were generated by visually determining the disparity based on the person's position relative to the ground truth disparity range map as shown in Figure IV.13. Experimental disparities were generated using the disparity voting algorithm with the disparity of each person determined from disparity values in the head region. A moving average of 150 ms was



(a) Frame 0 (b) Frame 20

(c) Frame 40 (d) Frame 60

(e) Frame 80 (f) Frame 100

(g) Frame 120 (h) Frame 140

Figure IV.11: Example Input Sequence for Multiperson Tracking Experiments. Notice occlusions, scale, appearance and disparity variations.

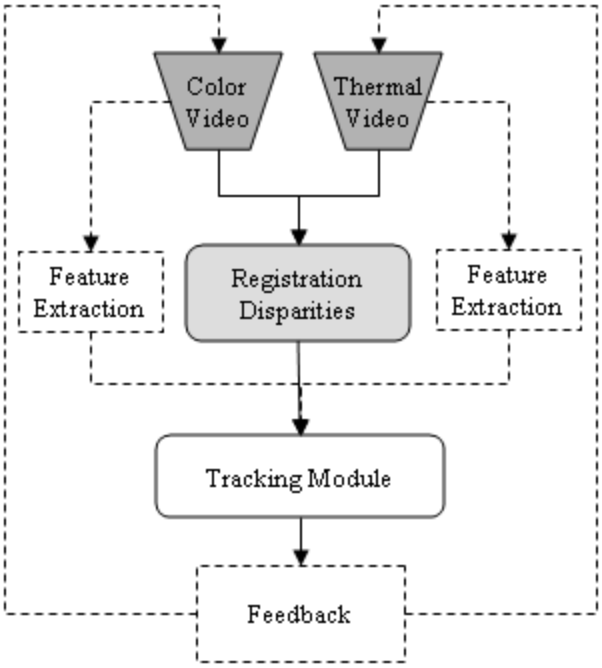
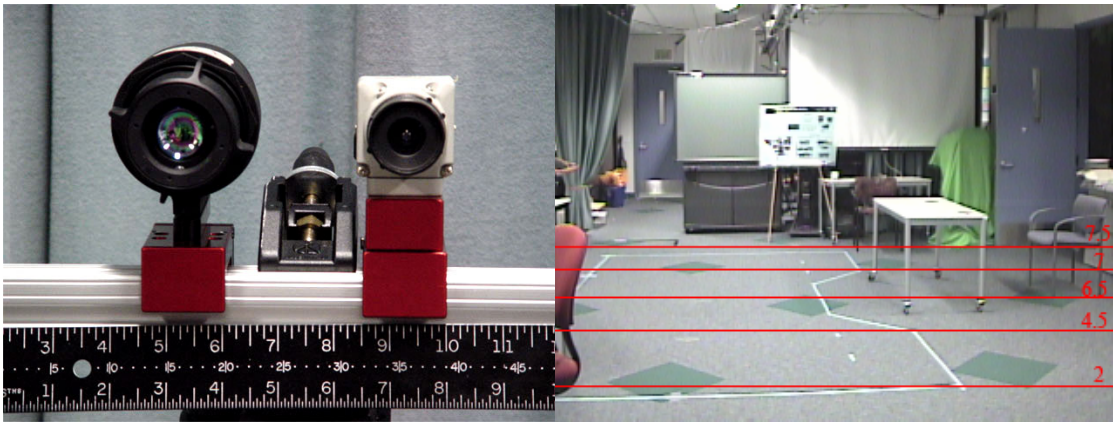


Figure IV.12: Algorithmic Flowchart for Multiperson Tracking



(a) Multimodal Stereo Rig (b) Disparity Range for Testbed

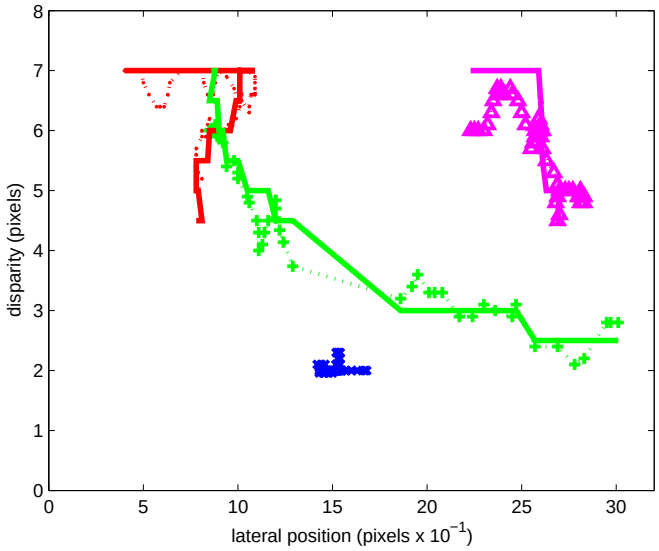
Figure IV.13: (a) Variable Baseline Multimodal Stereo Rig (b) Experimentally Determined Disparity Range for Testbed. The disparities were computed by determining the disparities for a single person standing at predetermined points in the imaged scene.

used to smooth instantaneous disparity estimates.

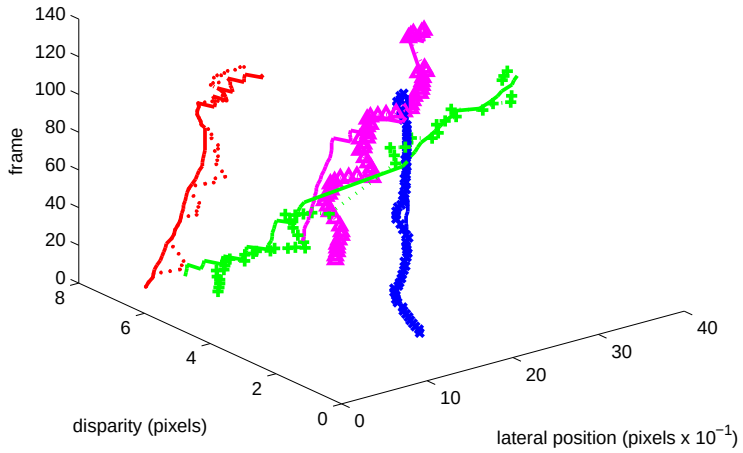
Figure IV.14 shows the track patterns and ground truth for the example sequence in Figure IV.11. The ground truth is plotted in solid colors for each person in the sequence, while the disparity estimates from the disparity voting algorithm are shown in corresponding colored symbols with dotted lines connecting the estimates. Figure IV.14(a) is a representation of the tracks, illustrating a "plan-view"-like representation of the movements and disparity changes of the people in the testbed. Figure IV.14(b) shows a time varying version of the same data, with the frame number plotted in the third dimension.

The plots in Figure IV.14 show that the disparities generated from the disparity voting registration reasonably follow the ground truth tracks. As the green tracked person moves behind and becomes occluded by the blue tracked person, we see that the disparities generated when he re-emerges from the occlusion are in line with the ground truth disparities and can be used to re-associate the track after the occlusion.

Errors from ground truth are particularly apparent when people are further from the camera. This is because of the non-linearity of the disparity distribution. There are more distinct disparities nearer to the camera. As you move deeper in the scene in Figure IV.13, the change in disparity for the same change in distance is much less. At these distances, errors of even one disparity shift are very pronounced. Conventional stereo algorithms typically used approaches that give subpixel accuracy, but the current implementation of our disparity voting algorithm only gives pixel level disparity shifts. While this may be acceptable for registration alignment, refinement steps are necessary to make disparity a more robust tracking feature. Approaches that use multiple primitives [24], such as edges, shapes, and silhouettes, etc., could be used to augment the accuracy of the disparity voting algorithm. Additionally, using multiple tracking features could provide additional measurements that can be used to boost the association accuracy.



(a) Track Patterns and Ground Truth for Four Person Tracking Experiment



(b) Time Varying Track Patterns and Ground Truth for Four Person Tracking Experiment

Figure IV.14: Tracking results showing close correlation between ground truth (in solid colors) and disparity tracked estimates (in dotted colors). Each color shows the path of one person in the sequence.

IV.H Summary

Multimodal imagery applications for human analysis span a variety of application domains, including medical [1], in-vehicle safety systems [2] and person detection [37]. Often, the registration algorithms these types of systems employ do not operate on data that has multiple objects and multiple depths that are significant relative to their distance from the camera. It is in this realm, including close-range surveillance [30] and pedestrian detection applications [31], that we believe disparity voting registration techniques and corresponding tracking algorithms will prove useful.

In this chapter we analyzed our method for registering multimodal images with occluding objects in the scene. By using the disparity voting approach, an analysis of over 2000 frames yielded a registration success rate of over 97%, with a 96% success rate when considering only occlusion examples. We have also shown relatively high success in outdoor environments when good segmentation is more difficult to achieve. Additionally, ground truth accuracy evaluations illustrate how the disparity voting algorithm provides accurate registration for multiple people in scenes with occlusion. Comparative studies show the improvements upon the accuracy and robustness of previous bounding box techniques in both a quantitative and qualitative manner. We have presented a framework for tracking and have shown promising experimental studies that suggest that disparity voting results can be used as a feature that will allow for the differentiation of people in a scene and give accurate tracking associations in complex scenes with multiple people and occlusions.

The text of this chapter, in part, is a reprint of the material as it appears in: Stephen J. Krotosky and Mohan M. Trivedi, "Mutual Information Based Registration of Multimodal Stereo Videos for Person Tracking" in *Computer Vision and Image Understanding, Special Issue on Advances in Vision Algorithms and Systems Beyond the Visible Spectrum*, Vol. 106, Issues 2-3, May-June 2007 and Stephen J. Krotosky and Mohan M. Trivedi, "On Color, Infrared and Multimodal Stereo Approaches to Pedestrian Detection", *IEEE Trans. On Intelligent Transportation Systems*, in press. I was the

primary researcher of the cited material and the co-author listed in these publications directed and supervised the research which forms a basis for this chapter.

Chapter V

Comparative Evaluation of Information Content in Color and Infrared Imagery: In Vehicle Pedestrian Detection

V.A Introduction

Pedestrian safety is a problem of global significance. Of the 1.17 million yearly worldwide traffic fatalities, 65% are pedestrian related [38]. In fully industrialized nations, pedestrian safety remains a high priority, with pedestrian fatalities accounting for 10.9% of all traffic deaths in the United States [39] and fatalities in Britain twice as likely for pedestrians than vehicle occupants [40]. In rapidly industrializing countries, pedestrian fatalities are overwhelmingly more costly in both proportion and sheer volume. Pedestrian and bicyclist fatalities in India were 80,000 in 2001, an estimated 60-80% of the total traffic deaths for that year [41]. Similarly, pedestrians and bicyclists accounted for 50% of all traffic related deaths in China in 1994 [42].

Naturally, such an important concern to public safety has received significant attention from all aspects of the research community. Specifically, ongoing computer vision research is making strides to detect and track pedestrians from both moving vehicles and the static transportation infrastructure. Typically, these approaches to pedestrian detection make use of visual or infrared imagery [43] in both monocular and stereo camera configurations.

The choice of visual or infrared imagery is significant, as each provides disparate, yet complementary information about a scene. Visual cameras capture the reflective light properties of objects in the scene, while infrared cameras are sensitive to the thermal emissivity properties of the same objects. Features extracted from each type of modality can be used to determine the presence of pedestrians in a scene. Additionally, by pairing each approach, their combination provides a level of features beyond what is readily obtained from the human visual system. Namely, the combination of visual and infrared imagery can provide the color, depth, motion, and thermal properties that can be used to more accurately detect, track and ensure pedestrian safety. Additionally, multiple camera systems have been incorporated into pedestrian detection approaches. The use of two or more cameras allows for the accurate depth estimates crucial to the task of pedestrian detection and collision mitigation.

This chapter presents research toward the development of a multimodal, multi-perspective system that can extract the depth and features necessary for robust pedestrian detection. We design a four camera experimental testbed consisting of two color and two infrared cameras for capturing and analyzing the various configuration permutations for pedestrian detection. Using this testbed, we perform comparative experiments of stereo-based detection approaches using unimodal color and infrared imagery and demonstrate the high obstacle detection rate achievable with stereo imagery. From these comparative experiments, we provide a detailed analysis of the features and properties of color and infrared imagery that are used to classify detected obstacles into pedestrian regions.

This analysis leads to our proposal of a multimodal trifocal framework consisting of a stereo pair of color cameras coupled with a single infrared camera. Using a calibrated three camera setup allows for accurate and robust registration of color, disparity and infrared features using the properties of the trifocal tensor. Under this framework, we demonstrate that the combination of color, disparity and infrared information can yield significant gains in pedestrian detection compared to detectors trained on only unimodal or stereo features.

V.B Related Research

Our focus on pedestrian detection is concerned with the methodologies and challenges of conventional camera systems. Specifically, we will review studies that utilize color and infrared imagery in single and multicamera configurations. For a more comprehensive review of computer vision based approaches to pedestrian detection, we refer the reader to a recent survey paper by Gandhi and Trivedi [44].

Single camera approaches were initially investigated to identify and localize pedestrians in a scene. To find pedestrians in crowded and varied scenes with changing backgrounds, typically a trained set of features that can identify pedestrian regions is extracted from the imagery. In color imagery, common features include Haar wavelet [45] or Gabor filter [46] responses, component-based gradient responses [47], image

contours with Mean Field models [48], Implicit Shape Models [49] and local receptive fields [50].

Similarly, features also need to be extracted from monocular infrared camera approaches. Typically the features extracted from infrared imagery are selected for their relation to the unique thermal signature of humans that enables straightforward segmentation. Such features that attempt to model this property include thermal hotspots [51], body model templates [52], shape independent multidimensional histograms, inertial and contrast base features [53] and Histograms of Oriented Gradients [54].

The features extracted from monocular imagery are then typically used in a classification scheme using many positive and negative examples of pedestrians. The most common approach to classification is to use a support vector machine (SVM) [45, 47,48,50,51,54,55]. Additional approaches to classification include template matching [52,56], convolutional neural networks [57] and Chamfer distance matching [49].

Despite the success of pedestrian detection approaches in monocular imagery, a single camera approach still has difficulty in one area critical to a fully realized pedestrian detection system: accurate and reliable depth estimates. To achieve this, a multi-camera systems is necessary, typically arranged in a stereo vision configuration. Visible light stereo systems [58–60] utilized the properties of dense stereo matching to robustly identify candidate pedestrian regions and accurately determine their distance from the camera setup. Infrared stereo camera systems have also followed that combine the benefits of infrared features with the powerful depth estimation inherent in stereo vision [56,61]. Additionally, a four camera system combining the separate approaches of color stereo and infrared stereo systems has been investigated [16]. In typical stereo camera systems for pedestrian detection, depth estimates are used to obtain a set of obstacle regions that are then further analyzed using monocular image features for pedestrian candidate generation.

V.C Stereo-based Pedestrian Detection

A fundamental step to analyzing pedestrians is to detect obstacles in the scene and localize their position in 3D space. A stereo camera setup is often used to obtain depth estimates of objects in the scene. A wide variety of algorithms can be implemented to obtain dense and accurate depth estimates from matching correspondences in calibrated and rectified stereo pairs [23].

The disparity images derived from stereo analysis are then used to generate a list of candidate pedestrian regions in the scene. We adapt a classical approach to obstacle detection in stereo imagery proposed by Labayrade *et al.* [62] that utilizes the concept of *v-disparity* to identify potential obstacles in the scene. Essentially, v-disparity is a histogram of the disparity image that counts the occurrence of disparity values for each row in the image. This histogram is very useful when the camera is relatively parallel to the imaged scene and objects appear at distinct planes in the disparity domain as the v-disparity information can be used to detect the ground plane in the scene and isolate regions that contain obstacles. Variations of this approach to detecting objects in stereo imagery have been implemented in [16, 59, 60].

V.C.1 Disparity-based Obstacle Detection

Our goal is to provide a framework for a comparative analysis of color and infrared stereo imagery for pedestrian detection. We have chosen to use the relatively simple v-disparity approach to obstacle detection so that it can be implemented for both color and infrared stereo imagery without modification or specialization. To that end, we examine the ability of each to generate stereo disparities and determine obstacle areas in the scene. This comparison of low-level detection accuracy will then lead to an evaluation of each camera type's potential for higher level obstacle classification and analysis. Figure V.1 shows a flowchart of the obstacle detection algorithm.

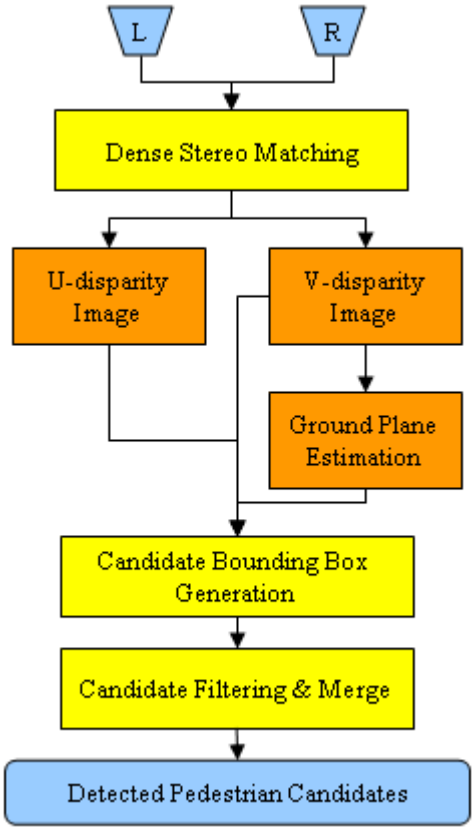
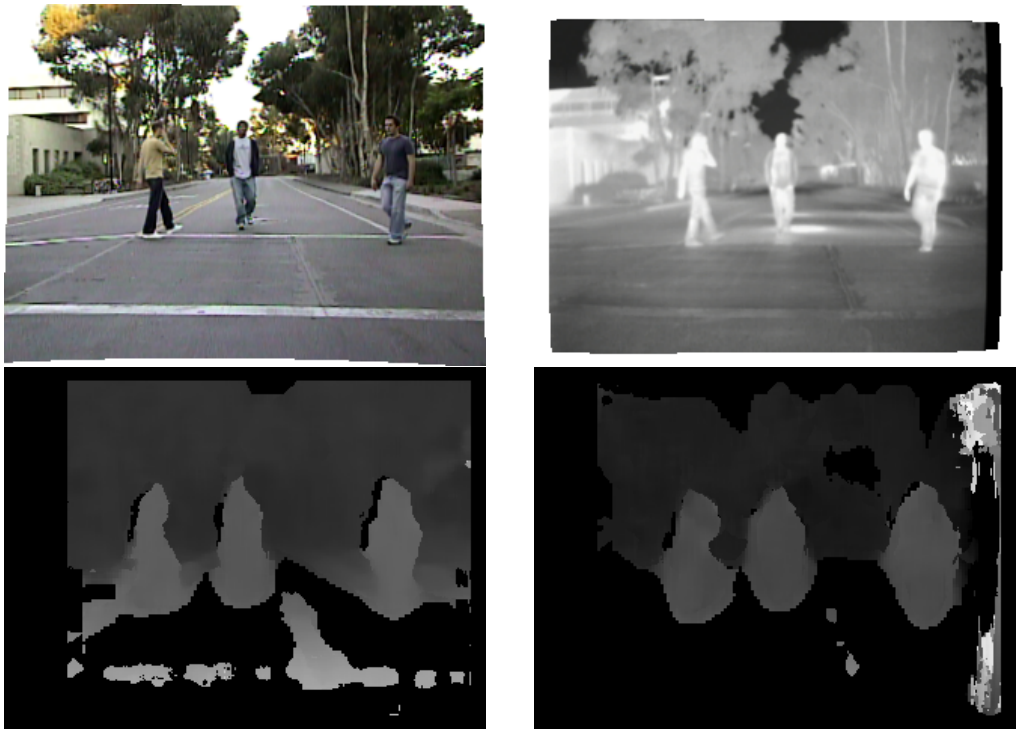


Figure V.1: Flowchart of stereo disparity-based obstacle detection algorithm.



(a) Color (b) Infrared

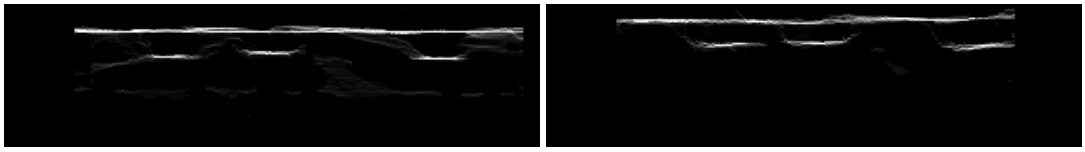
Figure V.2: Example disparity images from color and infrared stereo input images.

Dense Stereo Matching

As a first step, it is necessary to perform dense stereo matching to yield disparity estimates of the imaged scene. We elect to use the correspondence matching algorithm developed by Konolige [63] for its ease of use and reliable disparity generation with both color and infrared stereo imagery. Example disparity images generated using this approach are shown in Figure V.2.

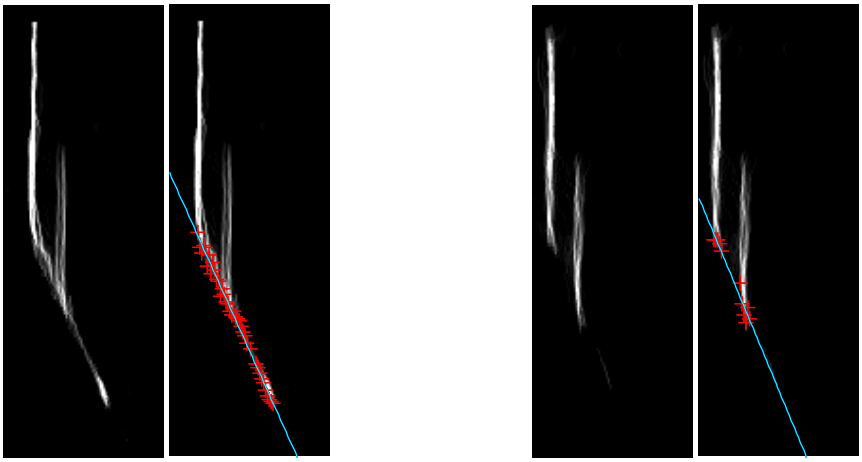
U- and V-Disparity Image Generation

The u- and v-disparity images are histograms that accumulate the number of pixels at a given disparity value, d, for each column or row in the image, respectively. For example, the v-disparity image is constructed so that for each row v in the disparity image D, the corresponding row in the v-disparity image is the histogram of those disparities in that row. The resulting v-disparity histogram image indicates the density of



(a) Color (b) Infrared

Figure V.3: Example u-disparity images from color and infrared stereo input images.



(a) Color w/ Ground Plane (b) Infrared w/ Ground Plane

Figure V.4: Example v-disparity images from color and infrared stereo input images along with the detected ground plane.

disparities for each image row v, while the u-disparity image shows the density of disparities for each image column u. Figure V.3 shows an example u-disparity image for color and infrared stereo imagery, and Figure V.4 shows the corresponding v-disparity images generated from the color- and infrared-based stereo disparity maps in Figure V.2.

Notice how the u-disparity images in Figure V.3 show three distinct horizontal regions of high disparity density corresponding to the three pedestrians in the scene. It is these regions we wish to detect in order to help build candidate pedestrian areas. The image spanning high density region at the top of the u-disparity image indicates the background disparities of the image and can be detected and filtered from processing. Similarly the v-disparity images in Figure V.4 show vertical peaks of high density for both the background plane and the range of disparities in D containing pedestrians. These regions will also need to be detected to generate pedestrian candidates. Addi-

tionally, there is a distinct downward sloping trend for the lowest image point for each disparity in the v-disparity image. It has been shown that this phenomenon can be used to estimate the ground plane of the image [62].

Ground Plane Estimation

To derive an estimate of the line indicating the ground plane, we must first extract candidate points on that line. For each column corresponding to a disparity d in the v-disparity image, we select the lowest pixel location whose value is above a given threshold as a candidate point in the ground plane. If there is no value that exceeds that threshold of a given disparity, then that disparity is omitted from the list of candidate points. Once the candidate points are obtained, the ground plane is estimated by fitting the candidate points to a line with a robust linear regression scheme that uses weighted least squares that iteratively reweights at each iteration using the bisquare weighting function. Figure V.4(b) and Figure V.4(d) show the v-disparity images for color and infrared stereo imagery with the candidate ground plane points in red and the fitted ground plane estimate plotted in cyan. Because we are using a dense stereo correspondence algorithm with robust point candidate generation and linear least squares fitting, we are able to reliably estimate the ground plane with both color and infrared stereo imagery.

Candidate Bounding Box Generation

Bounding box candidates can be extracted by first identifying regions-of-interest in the u- and v-disparity images. Regions in the u-disparity image can be extracted by scanning along the rows of the image, corresponding to a given disparity value. Regions are identified as continuous spans along the row where the histogram value in the u-disparity image exceeds a given threshold. Figure V.5(a)(b) shows the extracted regions in green on the u-disparity image. Regions are also extracted in the v-disparity image by scanning the columns corresponding to disparity values d and summing the histogram value above the ground plane in that column. If this sum is greater than a threshold, then the region-of-interest is selected that spans from the ground plane to a maximum height

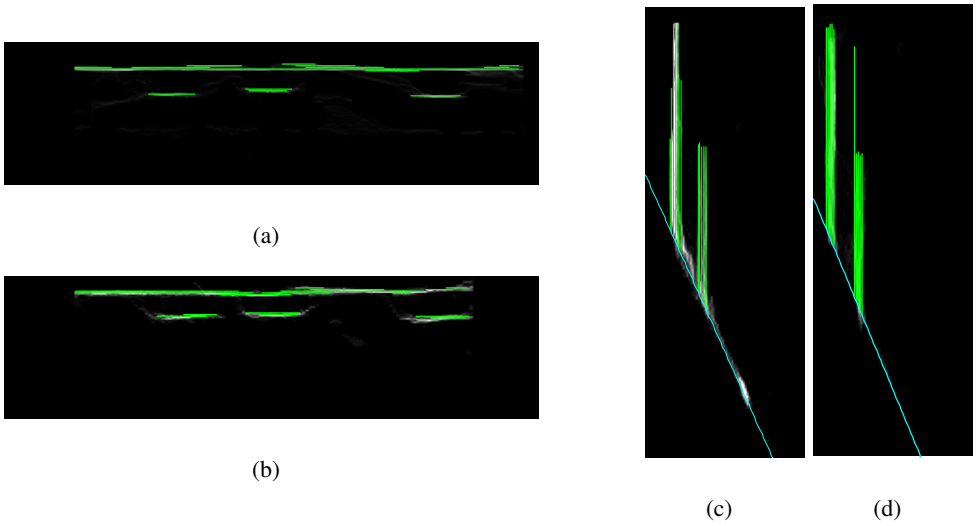


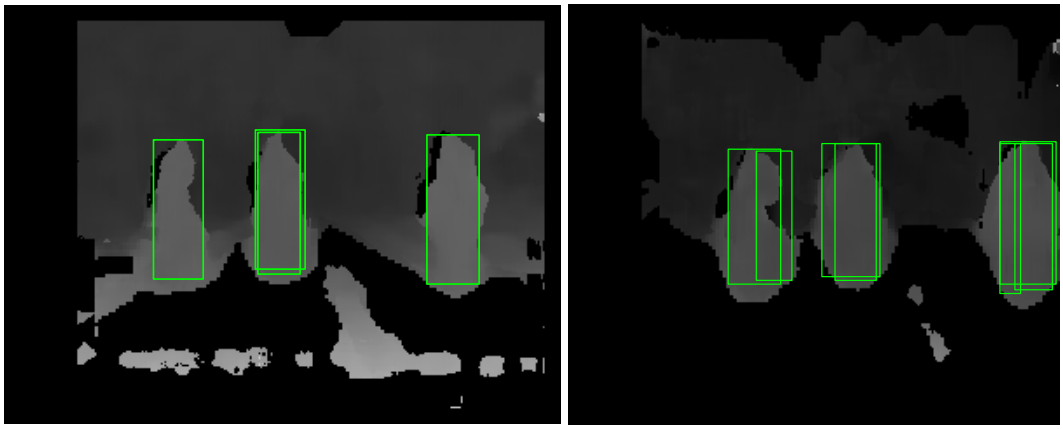
Figure V.5: Region-of-interest generation in u- and v-disparity images with color and infrared stereo input images. (a) Color u-disparity, (b) Infrared u-disparity, (c) Color v-disparity, (d) Infrared v-disparity

in the image where the histogram entry exceeds a given threshold. Figure V.5(c)(d) shows the extracted regions in green on the v-disparity image.

Candidate bounding boxes are then determined by associating the regions-of-interest in the u- and v-disparity images based on their disparity values. For a given disparity d, the width of the bounding boxes at that disparity are determined by the regions found in the u-disparity image and the height is correspondingly derived from the regions in the v-disparity image. Bounding boxes associated with the background regions that are obviously too large are removed. The resulting bounding box candidates are shown in green in Figure V.6.

Candidate Filtering and Merging

As shown in Figure V.6, there are often multiple overlapping candidate bounding boxes generated in the previous step. This usually arises because the disparities associated with a single pedestrian span a range of multiple values, especially if the pedestrian is closer to the camera. We merge overlapping bounding box candidates if their overlap is significant and the disparities associated with the bounding boxes are



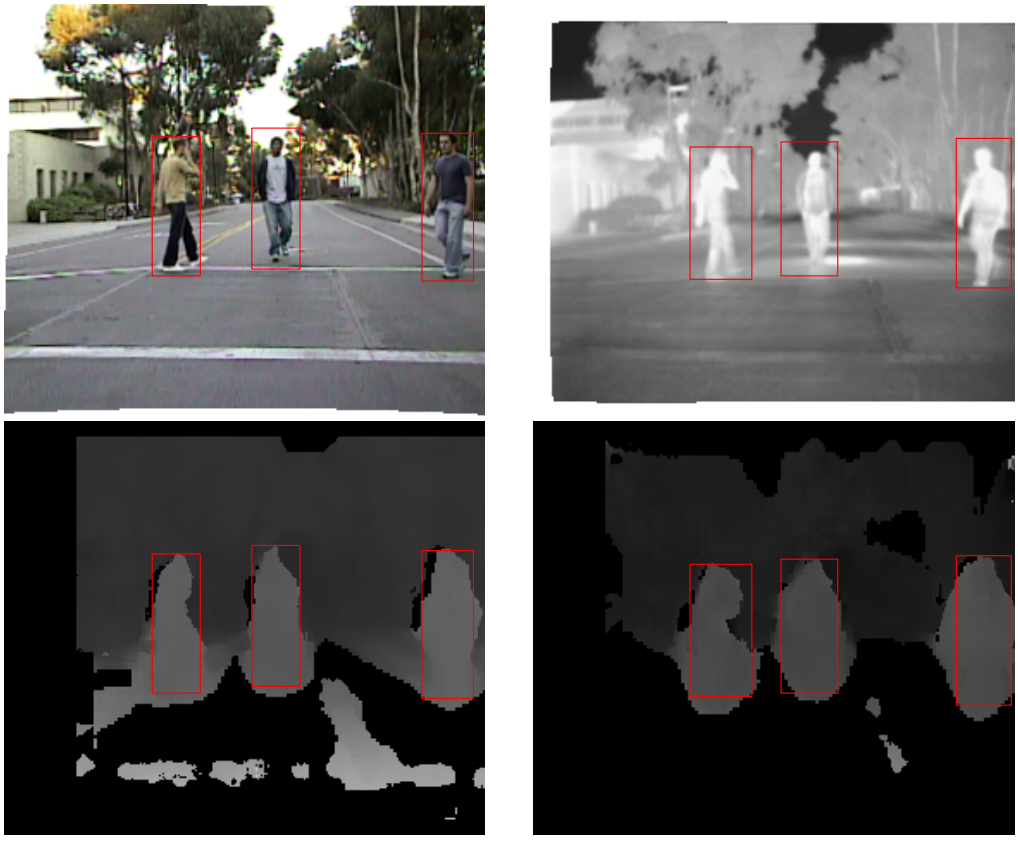
(a) Color (b) Infrared

Figure V.6: Example bounding box candidates with color and infrared stereo input images.

close. The resulting final selection of pedestrian candidate bounding boxes is shown in Figure V.7. Notice how the multiple bounding box candidates have merged into three appropriate bounding boxes associated with the correct pedestrians in the scene.

V.C.2 Experimental Framework and Testbed

We establish a framework for experimenting and analyzing pedestrian detection approaches for color and infrared stereo imagery. This framework needs to facilitate a direct, side-by-side comparison of the data coming from color and infrared stereo imagery. To that end, we have designed a custom rig consisting of a matched color stereo pair and a matched infrared stereo pair. The two pairs have been arranged so that their imaged scenes are as consistent as possible. The two pairs have identical baselines and the corresponding cameras in the color and infrared pairs are positioned as close as possible so as to maintain the same approximate fields of view. Additionally, lenses for the color cameras were selected to best match the fixed zoom of the infrared cameras. All four cameras are arranged in a single row and care was taken in aligning the pitch, roll and yaw of the cameras to maximize the similarity in field of view. Calibration data was obtained by placing a standard checkerboard pattern in front of the rig and illuminating



(a) Color (b) Infrared

Figure V.7: Example of the final selection of pedestrian candidates after bounding box merging with color and infrared stereo input images.

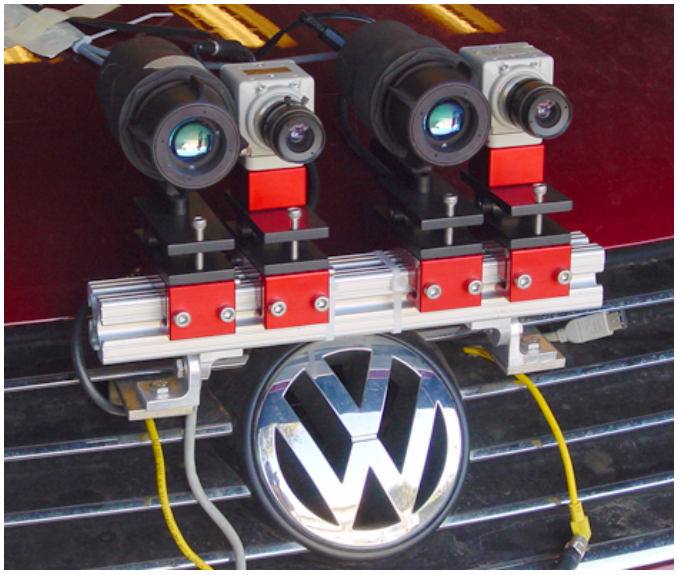


Figure V.8: Experimental testbed: Two color cameras and two infrared cameras arranged in stereo pairs and mounted to the front of the LISA-P testbed.

it with high intensity halogen bulbs so the checks could be viewed in both the color and infrared imagery. Once obtained, standard calibration techniques are used to obtain the intrinsic and extrinsic parameters of the cameras in the rig.

Once aligned, the rig was mounted to the grill of the LISA-P testbed described in Trivedi *et al.* [2] and [64]. The LISA-P is a Volkswagen Passat equipped with the computing, power, and cabling requirements necessary to synchronously capture and save the four simultaneous camera streams of our custom rig. Figure VI.8 shows the four camera rig properly arranged and mounted on the LISA-P. The two color cameras are gray while the infrared cameras are black.

V.C.3 Experimental Analysis of Disparity-based Obstacle Detection in Color and Infrared Stereo Imagery

Experiments were conducted where pedestrians would walk in front of the LISA-P testbed. The experiments included multiple pedestrians in the scene with varying degrees of depth, complexity and occlusion. The experimental data was captured simultaneously with the color and infrared stereo cameras to allow for direct compari-

son of the approaches. The captured data was analyzed using the disparity-based obstacle detection algorithm in Section V.C.1 and detection was determined successful if a bounding box correctly overlaid a corresponding pedestrian region. The bounding box must encapsulate the torso and head of a person and extend to the person's footage region. The bounding box must also not overestimate the person size by more than 10%. If two candidate bounding boxes associated with two separate pedestrians merged into a single bounding box after the merge process, we still consider the detection correct, yet note it as a "merge error" (Figure V.9). We reason that errors associated with a lack of sophistication of our chosen merging algorithm should not adversely affect the detection rate, as our desire is to evaluate the effectiveness of color and infrared stereo disparities to identify pedestrian areas and not the robustness of the merging procedure. This is also a fair assessment when using pedestrian detection for collision mitigation, as finding all the critical areas in the scene is given priority over discerning merged bounding boxes. Therefore, false negatives were counted only when a bounding box does not properly identify a pedestrian region (Figure V.10) and false positives were counted when a bounding box enclosed an area where no pedestrian existed. Still, had we incorporated the merge errors, the total detection rate would decrease by only 1% for color and 1.4% for infrared analysis. Table V.1 shows the compiled results of the comparative experiments and Figure V.11 shows additional examples for both color and infrared stereo inputs.

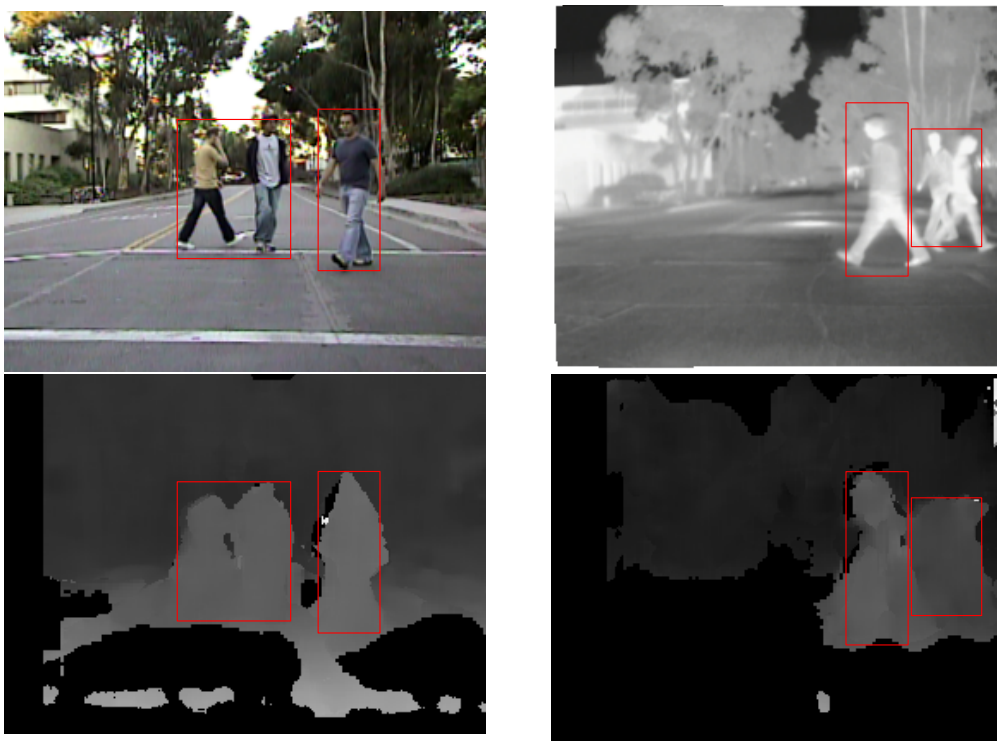
V.D Analysis of Stereo-based Pedestrian Detection

Our comparative experiments in Section V.C with stereo-based pedestrian detection for color and infrared imagery indicate a very high level of detection accuracy and low false positive rate in the both modalities. However, a deeper analysis of the experiments is necessary to truly understand and evaluate the success of these experiments.

We first note that although the analysis was performed on synchronously cap-

Table V.1: Results of experimental comparison between color and infrared stereo imagery for disparity-based obstacle detection.

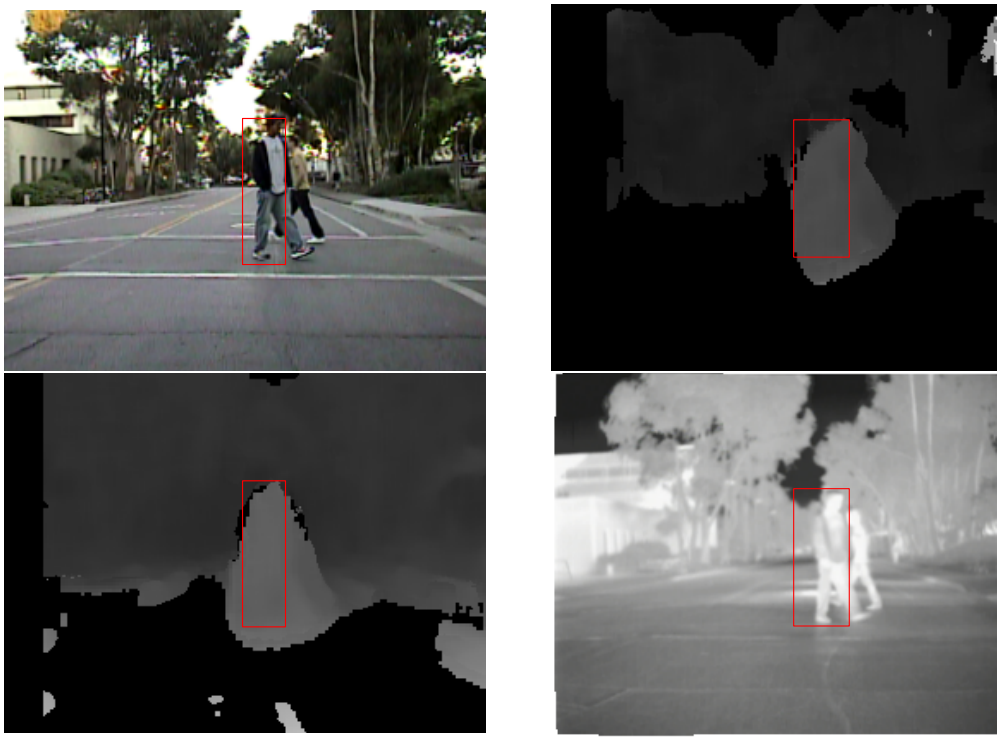
Modality	# Peds in Frame	Peds Correct	% Correct	False Positives	Merge Errors
Color	1	758 / 758	100.0%	0	0
	2	2376 / 2388	99.5%	2	7
	3	1525 / 1526	99.9%	0	35
	4	377 / 380	99.2%	1	6
	Total	5036 / 5052	99.6%	3	48
Infrared	1	880 / 899	97.9%	1	0
	2	2257 / 2287	98.7%	4	14
	3	1231 / 1244	98.9%	0	43
	4	123 / 124	99.2%	1	10
	Total	4491 / 4554	98.6%	6	67



(a) Color

(b) Infrared

Figure V.9: Example of merged pedestrian candidates with color and infrared stereo input images.



(a) Color

(b) Infrared

Figure V.10: Example of missed pedestrian candidates with color and infrared stereo input images.

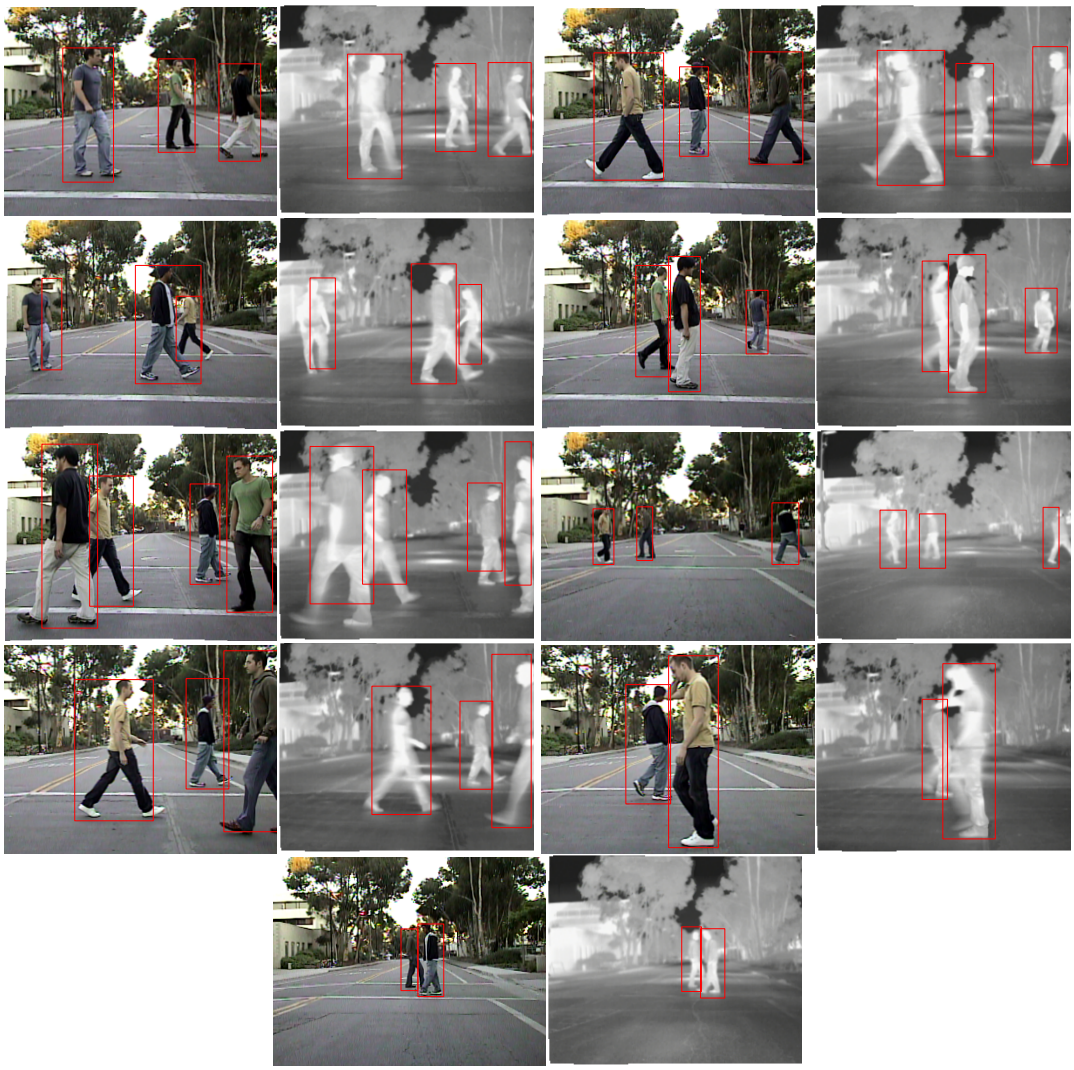


Figure V.11: Example of the final selection of pedestrian candidates with color and infrared stereo input images.

tured video, the pedestrian counts in Table V.1 differ for color and infrared imagery. The difference arises from the slight displacement in position of the color and infrared stereo cameras as well as the slightly smaller field of view of the infrared cameras. Only pedestrians that are fully visible in the image are considered, so there are many cases where a pedestrian is visible in the color image, but not seen in the infrared image. At 30fps, these instances quickly add up when comparing detections on a frame-by-frame basis. However, we feel that given the high number of examples, the detection rates can be compared even if the actual tallies differ in the color and infrared imagery.

The experiments yielded such a high rate of detection accuracy because our analysis equated low level obstacle detection with the higher level analysis of pedestrian determination. That is to say, since the experiments did not include non-pedestrian obstacles, such as other vehicles or bicyclists, a detection of any obstacle region is assumed to be a pedestrian. For the scope of our experiments, this sort of assumption is appropriate, as we are interested in evaluating the ability of color and infrared dense stereo correspondences to be used in low level pedestrian detection. In that respect, our experiments demonstrate that color and infrared stereo disparities both achieve high rates of low level obstacle detection, an imperative first step towards robust pedestrian detection and collision mitigation.

However, in real world driving scenarios, low level obstacle detection, while an imperative initial step, is not sufficient for pedestrian detection. Detected obstacles can include a variety of objects found in common driving scenes other than pedestrians, such as parked and moving vehicles, trees, buildings, parking meters and other spurious candidates in the scene. Additional processing is necessary to filter the detected obstacles into appropriate pedestrian and non-pedestrian regions.

In the disparity image domain, it is possible to filter some of the detected obstacles based on the bounding box features of typical pedestrian obstacles (e.g. Bertozzi *et al.* [16]). Bounds on pedestrian bounding box features such as size, disparity and aspect ratio can be learned or heuristically selected to filter out bounding boxes associated with other objects in the scene. However, the success of such filtering techniques can

prove unreliable, as it will not filter non-pedestrian bounding boxes that fall within the selected bounds of pedestrian candidates. Additionally, the selection of appropriately robust bounds is a challenging task, as bounding box sizes can vary with pedestrian pose and disparity fidelity. To achieve more reliable detection of pedestrian candidates, it is necessary to analyze other features of the imagery in addition to the stereo disparity estimates.

As mentioned in Section V.B, features that have been used for color image-based pedestrian classification include Haar wavelet responses [45], Gabor filter responses [46], Sobel edge responses [59], Implicit Shape Models with Chamfer distance matching [49], image contours with Mean Field models [48], and local receptive fields for support vector machine classification [50].

Infrared features for classification typically include features that identify the specific thermal characteristics of the scene, including hotspots [51], warm element and head template matching [56], body model templates [52], shape independent multidimensional histograms, inertial and contrast base features [53] and Histograms of Oriented Gradients [54].

Obstacle detection using stereo disparities derived from color or infrared imagery is highly accurate with low false positive rates. However, this level of detection is still too primitive to be used for real world pedestrian detection as it can include obstacles not associated with pedestrians or other critical regions. To supplement and filter these obstacle candidates, specific features of color or infrared imagery can be extracted and analyzed to determine the true pedestrian regions in the scene. Although both color and infrared imagery have been used to identify pedestrians in a scene, it is unclear which camera system is preferred. Because of the underlying differences in the physical processes that give rise to color and infrared imagery, the two modalities yield disparate information about the scene.

While a justification can be made for the use of either approach, a more interesting proposition would be to use both modalities in concert to obtain all sets of available features in color and thermal imagery. Naturally a detection architecture that

incorporates more features has a higher potential for detection accuracy than one with a lesser feature set. For example, the thermal "hotspots" of humans that often make pedestrians easily segmentable can be used together with the fine level of color image detail that has proven useful for tasks as challenging as detecting articulated poses for classifying human interactions [65].

Although it is possible to incorporate the advantages of stereo color and infrared analyses by separately combining the two camera systems and pedestrian detection [16], it is costly and cumbersome to incorporate a four camera solution from both a computational and vehicle integration standpoint. A more economical and desirable solution would be to combine the benefits of color, disparity and infrared imagery in an integrated detection framework. In Section V.E, we propose a multimodal trifocal framework consisting of a stereo pair of color cameras coupled with a single infrared camera. Using a calibrated three camera setup allows for accurate and robust registration of color, disparity and infrared features using the properties of the trifocal tensor. We use this robust registration to design a pedestrian detector that integrates color, disparity and infrared features to yield increased detection rates over detectors using unimodal or stereo-based features.

V.E Multimodal Trifocal Framework for Pedestrian Detection

The benefits of color, disparity and infrared image features can be incorporated using a three camera approach consisting of a standard color stereo rig paired with a single infrared camera. This trifocal framework, illustrated in Figure V.12, allows disparity estimates obtained through dense color stereo correspondence matching to register corresponding image pixels in the infrared imagery. This can be done quickly and efficiently by using the trifocal tensor – the set of matrices that relates the correspondences between the three images.

The trifocal tensor can be estimated by minimizing the algebraic error of point correspondences as described in Hartley and Zisserman [8]. These point correspon-

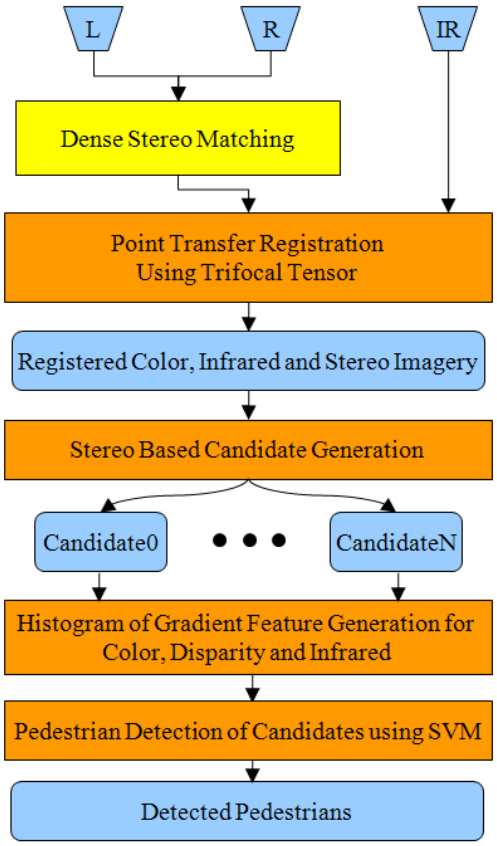
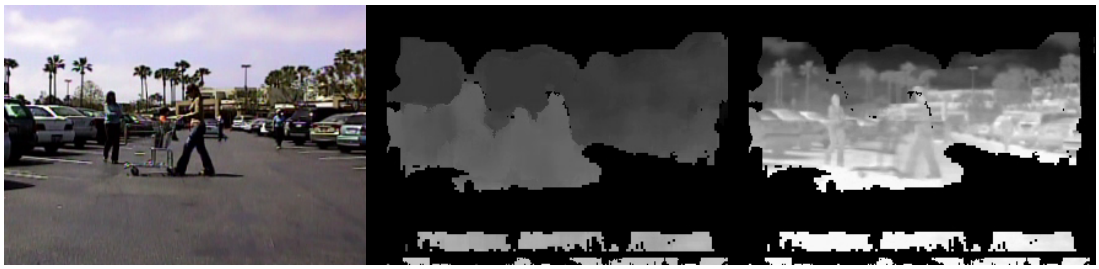


Figure V.12: Flowchart of trifocal tensor approach to pedestrian detection for color stereo and infrared framework.



(a) Color (b) Disparity (c) Aligned Infrared

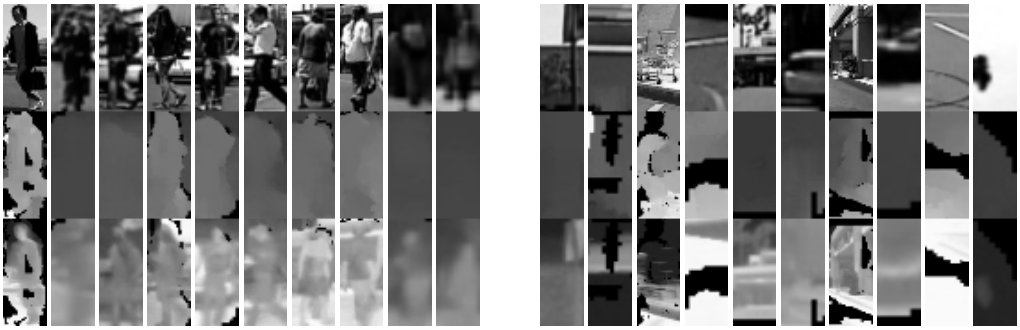
Figure V.13: Example of registered color, disparity and infrared imagery using trifocal tensor.

dences can be obtained for all three images using the same calibration board and techniques used for stereo calibration. We obtain images where the calibration is visible in each of the three views of the multimodal trifocal framework. While only seven point-point-point correspondences are required to compute the trifocal tensor, in practice, as with stereo calibration, many more correspondences are used to smooth errors in the estimates. The resulting trifocal tensor is written as $\mathcal{T} = [T_1, T_2, T_3]$, where T_i is a 3x3 matrix for the i^{th} image in the set. From this tensor notation, standard two-view geometry parameters, such as fundamental matrices F, epipoles $\boldsymbol{e}$ and projection matrices P can be evaluated.

Additionally, using the trifocal tensor notation, given a point correspondence $x' \leftrightarrow x''$, we can estimate the point transfer to the third image point x using the following method:

$$[x']_\times (\sum_i x^i T_i)[x'']_\times = 0_{3\times 3} \tag{V.1}$$

Dense stereo correspondence matching gives $x' \leftrightarrow x''$ correspondences for a large portion of the scene and the trifocal tensor technique is able to register infrared pixels at these valid disparity regions. The resulting point transfers are all aligned to the color reference image. An example set of aligned images is shown in Figure VI.4.



(a) Positive Samples (b) Negative Samples

Figure V.14: Selection of positive and negative samples used for training pedestrian detectors. Each sample consists of color, disparity and infrared images.

V.E.1 Experimental Evaluation of Pedestrian Detection using Color, Disparity and Infrared Image Features

To determine the effect of using multimodal features for pedestrian detection, we use the trifocal framework to register the color, disparity and infrared imagery into a single five-channel multispectral image. The single alignment of all image data allows for fast and easy generation of positive and negative samples. Using the same testbed from Sec. VI.E.1, we are able to design pedestrian detectors that make use of various combinations of color, disparity and image features. To train the detectors, positive pedestrian samples are manually annotated in the five-channel multispectral image. For each positive sample, 10 negative samples are generated by moving the positive bounding box to a random non-overlapping position in the test image. All samples are resized to a common size (24x60 pixels) as shown in Figure V.14.

For each sample, features need to be extracted that can be used for training the detectors. We elect to use Histograms of Oriented Gradients (HOG) similar to those proposed by Dalal and Triggs [66]. For each of the color, disparity and infrared images, we compute a XxYxΘ element histogram, where X, Y and Θ are the number of histogram bins in width, height and gradient orientation, respectively. For our experiments, we use 4x4x8 element histograms, resulting in a 128 element feature vector for each image in the sample. The selection of this feature descriptor is based on the notion that gradient

Table V.2: Pedestrian detection rate for 5% false positive rate.

Color	86.74%
Disparity	71.45%
Infrared	74.10%
Color+Disparity	86.58%
Disparity+Infrared	83.31%
Color+Infrared	84.40%
Color+Disparity+Infrared	91.89%

information will play a role in determining the presence of a pedestrian in a sample. While we do not claim that this is the best or optimal feature for detecting pedestrians in color, disparity or infrared imagery, HOG features have been shown to experimentally outperform similar features for detecting pedestrians [67] so we feel they are sufficient to evaluate how the combination of multispectral image features can improve detection accuracy.

Once the features have been computed, we train the pedestrian detector using a support vector machine (SVM) with a radial basis function as the kernel type [68]. We wish to use the SVM to derive a pedestrian detector for all combinations of color, disparity and infrared features. We train each SVM using 865 annotated positive samples (and 8650 negative samples) collected from video obtained while driving the LISA-P testbed in store parking lots and local roads in La Jolla, California. Similarly, to evaluate each detector, we use a test set of 641 positive samples and 6410 negative samples obtained from a different set of videos obtained while driving the LISA-P in La Jolla. Pedestrians in the training and testing sets range from approximately 3 to 30 meters from the vehicle. The resulting ROC curves for each detector are plotted in Figure V.15 and the detection rates given a 5% false positive rate are shown in Table V.2.

Clearly, the detector that uses the combination of color, disparity and infrared features performs better than all other detectors by a significant margin. By integrating the features, we exploit the complementary nature of multimodal imagery yielding more than 5% increase in detection for a fixed 5% false positive rate. This indicates that incorporating multimodal features increases the robustness of detecting pedestrians. Ad-

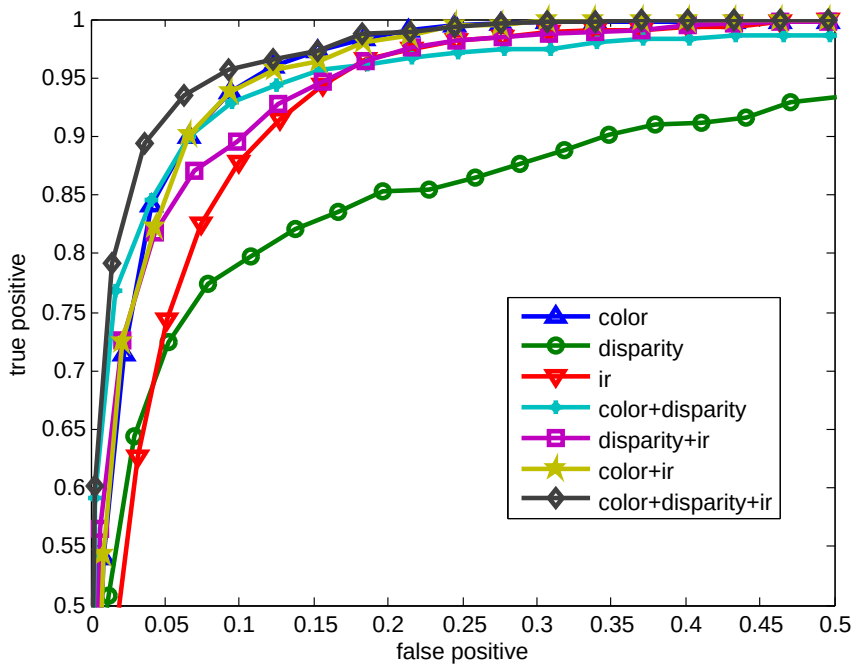


Figure V.15: ROC for pedestrian detection. The combination of color, disparity and infrared features performs the best.

ditionally, it can be anticipated that detection rates will be even higher in practice, as the negative samples were drawn from random portions of the image and likely would be filtered out by the initial object identification algorithm described in Section V.C.

We also note that the combinations of color+infrared and color+disparity do not outperform the detector trained on color alone. We suspect this is because the gradient based features used are not ideal for discriminating pedestrians from background in the low contrast disparity and infrared domains. This is evident when examining the performance of the detectors trained on only disparity or infrared images. It seems that given the relatively low number of positive samples, the addition of disparity or infrared adds more noise to the detector. It's all the more interesting then, that the color+disparity+infrared trained detector performs so well. The discriminant gains from combining all features greatly outweighs the noise added from using non-ideal gradient features. We anticipate even greater gains in detection accuracy could be achieved by combining each image spectrum using more discriminant features.

V.F Summary

The use of multimodal and multiperspective imagery has helped researchers take large steps towards achieving accurate and robust pedestrian detection. The depth estimates obtainable from vehicle mounted stereo imagery give a straightforward approach to extracting obstacle regions from the scene. We have outlined such an algorithm for obstacle detection in stereo imagery and have provided comparative experiments to gauge the detection rates achievable with color and infrared stereo imagery. Our analysis indicates that color and infrared-based stereo disparities are capable of highly accurate pedestrian detection ($> 98\%$) with low false positives ($\ll 1\%$).

Given the high detection rates obtainable from color and infrared stereo imagery, the selection of an appropriate camera system for pedestrian detection turns to the consideration of each modality's ability to further classify detected obstacles into pedestrian and non-pedestrian regions. Because the physical processes that give rise to color and thermal imagery are disparate, the extractable features from color and infrared imagery are also very different and largely unique to each modality. As previous approaches have demonstrated the usefulness of features from both color and infrared imagery for classifying pedestrian regions, we suggest that a complementary system that utilizes all the available features of color and infrared imagery is most desirable. Specifically, we propose a multimodal trifocal solution to obtain the color, depth and infrared features desirable for pedestrian detection.

The multimodal trifocal solution, consisting of a color stereo rig paired with a single infrared camera allows for accurate and robust registration of pixels in each image. Using the color stereo aspect, we can achieve the same high level of obstacle region identification as in the unimodal case. However, we demonstrate that integrating color, disparity and infrared features for training a pedestrian detector yields improved accuracy over detectors that utilize only unimodal or stereo features. From a cost-benefit perspective, we suggest that the multimodal trifocal framework is likely the best approach, as it can achieve the benefits of multimodality seen in higher camera solutions,

yet achieves robustness beyond what is capable in two camera cross-spectral solutions. Future areas for investigation include a more extensive evaluation of feature selection in color, disparity and infrared imagery. Additionally, an integrated object candidate generation and pedestrian detection algorithm using the multimodal trifocal framework would be useful for evaluating the robustness of pedestrian detection in various lighting and environmental conditions.

These multimodal and multiperspective approaches provide useful insight into the overall active safety paradigm. Pedestrian safety is just one of many aspects of the driving environment that needs to be monitored in order to enhance safety in the vehicle and surrounding areas [64]. The multimodal feature set extractable from the multimodal trifocal solution could provide for a robust and unified framework for detecting pedestrians and other obstacles in the vehicle surround [69]. Additionally these features can be used for higher level driver intent analysis such as lane changing [70], turning [55] and braking [71].

The text of this chapter, in part, is a reprint of the material as it appears in: Stephen J. Krotosky and Mohan M. Trivedi, "On Color, Infrared and Multimodal Stereo Approaches to Pedestrian Detection", *IEEE Trans. On Intelligent Transportation Systems*, in press. I was the primary researcher of the cited material and the co-author listed in this publication directed and supervised the research which forms a basis for this chapter.

Chapter VI

Comparative Evaluation of Information Content in Color and Infrared Imagery: In Vehicle Pedestrian Detection: Surveillance

VI.A Introduction

This chapter presents a methodology for analyzing multimodal and multiperspective systems for person surveillance. Using an experimental testbed consisting of two color and two infrared cameras, we can accurately register the color and infrared imagery for any general scene configuration so the scope of multispectral analysis can be expanded beyond the specialized long-range surveillance experiments of previous approaches to more general scene configurations common to unimodal approaches.

We design an algorithmic framework for detecting people in a scene that can be generalized to include color, infrared and/or disparity features. Using a combination of histogram of oriented gradient (HOG) features from the color and infrared domains, we train a support vector machine to detect people in the scene. Additionally we learn the relationship between person size and depth in the scene to create a disparity-based detector. We assume that the visual and disparity trained detectors can be treated independently and probabilistically combine their outputs to create an overall detection score.

Within this framework, we train person detectors using color stereo and infrared stereo features. We also analyze tetravision-based detectors that combine the detector outputs from separately trained color stereo and infrared stereo features. Additionally, we incorporate the trifocal tensor in order to combine the color and infrared features in a unified detection framework, doing so for both the color stereo + single infrared and infrared stereo + single color cases. We use these trained detectors for an experimental evaluation of video sequences captured with our designed test bed.

Our evaluation definitively demonstrates the performance gains achievable when using the trifocal framework to combine color and infrared features in a unified framework. Both of the trifocal setups outperform their unimodal equivalents, as well as the tetravision based analysis. Our experiments also demonstrate how the trained detector generalizes well to different scenes and can provide robust input to an additional tracking framework.

VI.B Related Research

Person analysis in multispectral and multiperspective imagery is a relatively new area of research in computer vision. Analysis that incorporates a comparison between color and infrared imagery for person analysis has been relatively sparse and limited in scope and generality.

Typical studies have looked at person detection by treating color and infrared separately. For example, Zhang *et al.* [67] compared different image features in color and infrared monocular imagery for training a support vector machine. However, no direct comparison of the detection rates of color and thermal imagery was presented. Ran *et al.* [72] also looked at separately using color and thermal imagery to detect periodic motion to indicate pedestrians in the scene. The main goal of these studies is to show the extensibility and adaptation of color image analysis techniques on infrared imagery.

Other studies have examined person detection as a fusion of color and infrared imagery. Davis and Sharma [3] have constructed a data set of color and infrared videos. The data set provides for a frame-by-frame comparison of the color and infrared imagery and also allows for the registration of the two videos as the view conforms to a planar homography assumption. This data set has allowed for the development of algorithms that combine the color and thermal imagery for improved background subtraction [73] [9] and person detection and tracking [74].

The planar homographic assumption is a convenient way to register the color and infrared imagery. However, the assumption is also severely limiting in the types of scenes that can be analyzed with multiperspective imagery. Because it is assumed that all objects can be aligned with a single planar homographic transformation, the scene must be in a special configuration to achieve this assumption. Typically this means that all objects are sufficiently far from that camera that they satisfy the infinite homographic assumption. While this provides a method of registration, other scenes where people can be at multiple distances from the camera, such as those commonly analyzed in monocular and stereo imagery cannot be analyzed in the planar homographic framework. Our

previous studies [20] have elucidated the ways that multispectral and multi-perspective data can be registered and have shown ways to register color and infrared imagery for any general scene configuration.

The most effective way to register color and thermal imagery for any general scene context is to incorporate stereo imagery whose depth estimates can account for the parallax inherent in any multiperspective scene. Bertozzi *et al.* [16] [56] has designed a four camera "tetravision" system to analyze people in color and infrared stereo imagery. Detection is performed separately in the color stereo and infrared stereo domains. The detection results are then fused by associating the detected bounding boxes from each modality based on their 3D location in the scene.

We have introduced a trifocal approach to person detection with color and thermal imagery [37]. By incorporating stereo depth estimates from a single modality we can register the second modality accurately using the trifocal tensor. This framework gives a method for combining the color and infrared features to design a unified multispectral classifier that can be used to improve the accuracy and robustness of unimodal detection frameworks.

VI.C Trifocal Tensor vs. Homography

The trifocal approach to combining color and infrared imagery allows us to compare a wider range of data than has previously been analyzed in the literature. Typical approaches that combine color and infrared for analyzing pedestrians focus on scenes where objects appear very far away from the camera, such as those from the IEEE OTCBVS WS Series Bench [3]. This allows for straightforward registration using a planar homography assumption, yet limits the depths of field that can be analyzed. This means analysis must be confined to a restricted plane-of-interest in the scene, or the cameras must be placed to ensure that all areas in the scene will comply with the homography.

Figure VI.1 shows the differences in fields-of-view between (a) typical planar

(a) Typical Data in Planar Homographic Framework

(b) Typical Data in our Trifocal Framework

Figure VI.1: Comparison of viable field of view for combining color and infrared imagery for (a) for planar homography, and (b) our trifocal approach.

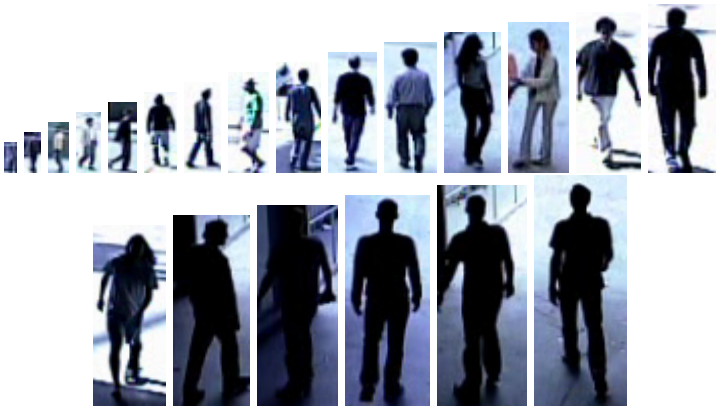


Figure VI.2: Range of scales at which people can be seen in trifocal framework.

homographic and (b) our trifocal approach to color and infrared analysis. Notice how people in the planar homographic imagery are all very far away from the camera and at a similar scale. This is a requirement of the planar homographic approach and puts severe limits on the types of scenes that can be fully analyzed within this framework. The typical data from the trifocal framework is much more general and complex. People can be at a broad range of scales and distances from the cameras. As long as depth estimates for an image region can be obtained, we can register the relevant pixels of objects at any general position in the scene.

Figure VI.2 illustrates the large range of scales that can be obtained in the trifocal framework. It is a challenge to design a person classifier that is able to handle such a broad range of scales as the extracted features need to be relatively invariant to these scale changes. The incorporation of this scale range also greatly increases the number of candidates to consider, thereby increasing the potential for false positives.

VI.D Algorithmic Framework

We wish to explore how the incorporation of color, infrared and disparity features affect the classification and false positive rates of a person detection system. To do so, we establish a framework for registering the multimodal imagery and extracting features from this imagery that can be used to learn to detect people in a scene. Figure VI.3

shows the algorithmic flow of this framework. We describe the details of our framework in the following sections.

VI.D.1 Image Registration with Trifocal Tensor

We use a three camera approach, consisting of a unimodal stereo pair (color or infrared) combined with a single camera of the second modality. We use the disparity estimates from the stereo imagery to register corresponding pixels in the third image with the trifocal tensor – the set of matrices relating the correspondences between the three images. The multimodal trifocal registration algorithm is identical to the one described in Section V.E. Figure VI.4 shows an example set of registered trifocal imagery for these experiments.

VI.D.2 Annotation

Now that we are able to accurately register all three modalities, we can extract positive and negative samples for classification. Positive samples need to be annotated from video sequences. Bounding boxes for all people in the scene were annotated. For consistency in classification, all bounding boxes maintain a 2:5 aspect ratio. Negative samples can then be generated by translating the corresponding bounding box for a person to a non-person region in the scene. Additional negative samples are generated by selecting smaller sub-regions of the selected pedestrian region. Although annotation needs to be done once for a single trifocal setup, it is necessary to repeat the annotation for both the color stereo and infrared stereo frameworks, as they have slightly different fields-of-view. Care was taken to ensure that samples were annotated from identical people at identical frames in both cases to limit variability in the training. Additionally, only non-occluded pedestrians were included in the training set. We expect the classifier will still be able to handle occlusion without explicitly training for it and our experimental evaluation will validate this assumption. Figure VI.5 shows example positive samples annotated in color and infrared reference imagery. For each sample, we can simultaneously extract the reference image patch, its disparity image and the reprojected

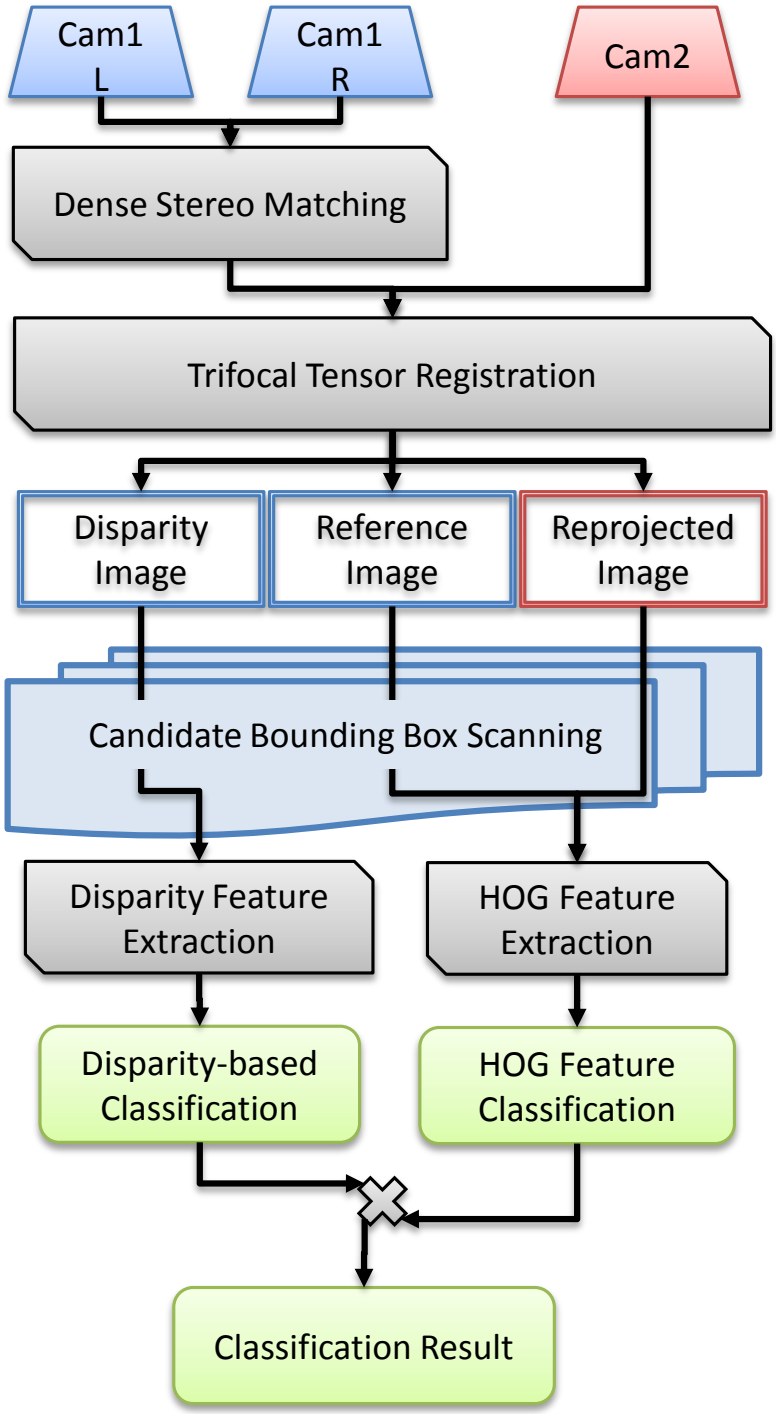


Figure VI.3: Algorithmic framework for person detection with color, infrared and disparity image features.

(a) Color Reference (b) Infrared Reference

(c) Color Disparity (d) Infrared Disparity

(e) Registered Infrared (f) Registered Color

Figure VI.4: Examples of using the trifocal tensor to register a third image to a stereo pair. The left column shows an infrared image registered to a color stereo pair and the right column shows a color image registered to an infrared stereo pair.

(a) Color Samples (b) Infrared Samples

Figure VI.5: Example positive samples of people extracted from (a) color stereo reference images and (b) infrared stereo reference images. The top row shows the reference sample, the middle row shows the disparity sample and the bottom row shows the re-projected image sample.

image data to create the combined sample triplet.

VI.D.3 Image Features

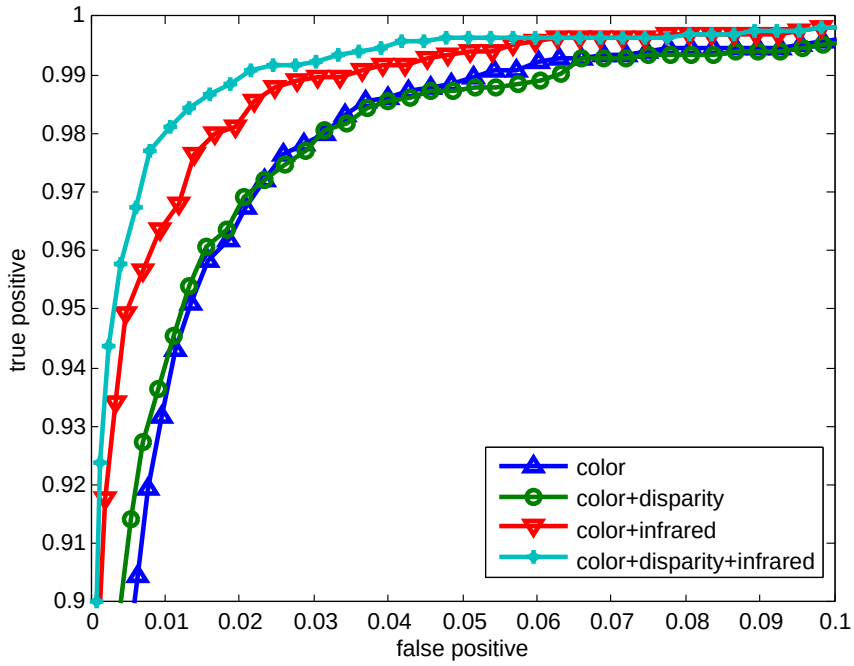
Once annotated, we must extract features that will be able to differentiate the positive and negative samples. For the color and infrared images, we elect to extract Histograms of Oriented Gradients (HOG) features similar to those proposed by Dalal and Triggs [66]. These features attempt to encode the relevance of edges in terms of their orientation and spatial position and have been increasingly utilized in many recent person classification publications [54] [67]. We resize each of the color and infrared image samples to a common size and compute an $X \times Y \times \Theta$ element histogram, where X, Y and Θ are the sizes of histogram bins in width, height and gradient orientation, respectively.

For the disparity image, we initially considered extracting HOG features as well. Our initial results indicated this was valid, as the ROC curves showed that the classifier trained on color, infrared and disparity HOG features outperformed those trained on just color and infrared. Figure VI.6 shows the ROC curves for classifiers trained on variations of color, infrared and disparity HOG features. Figure VI.6(a) shows the ROC curves for the color stereo reference and Figure VI.6(b) shows the ROC curve for the infrared color reference. The combination of color, infrared and disparity performs the highest when evaluating cropped samples in a cross-validation framework. This is a misleading result, though, as the ROC is constructed only by classifying annotated image patches. When the same classifier is applied to find people in novel images, the resulting regions include many false positives, often more than the number of people in the scene.

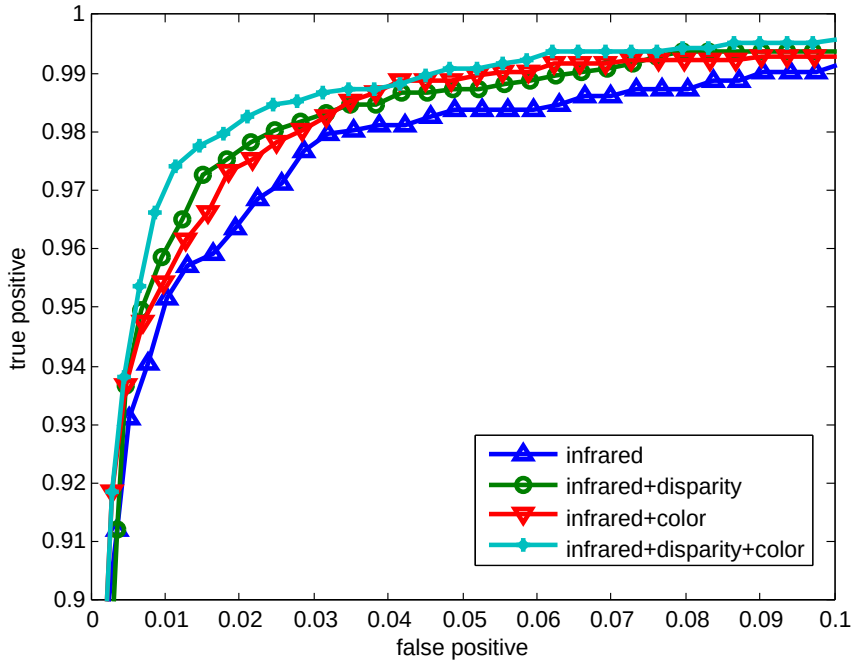
This performance drop-off is likely due to a combination of factors. First, it is likely that the HOG features are inappropriate for disparity imagery. They are designed to capture edge properties of an image patch, yet for many positive samples of people, there are few to no edges in the disparity image. This is especially true for people close to the background, where their difference in disparity from the background is small. While adding these HOG disparity features can provide some additional differentiability when classifying the carefully cropped and annotated image patches, the features actually give false positives when classifying novel images, especially at regions near true person regions.

To find an alternative, we further examine the disparity imagery to find features that help to differentiate people from the background and other objects in the scene. We notice that there is a linear correlation between the size of the bounding box that encloses the person and the median of the disparity inside that region. Figure VI.7 shows the relationship between the bounding box height and the median disparity and the least-squares linear fit of the data.

This line is parameterized as $Ax + By + C = 0$, where (A, B, C) are the parameters of the line, x is the image height and y is the median disparity. For a candidate



(a) Color Reference



(b) Infrared Reference

Figure VI.6: ROC curves showing the combination of color, disparity and infrared features when using HOG features for all modalities

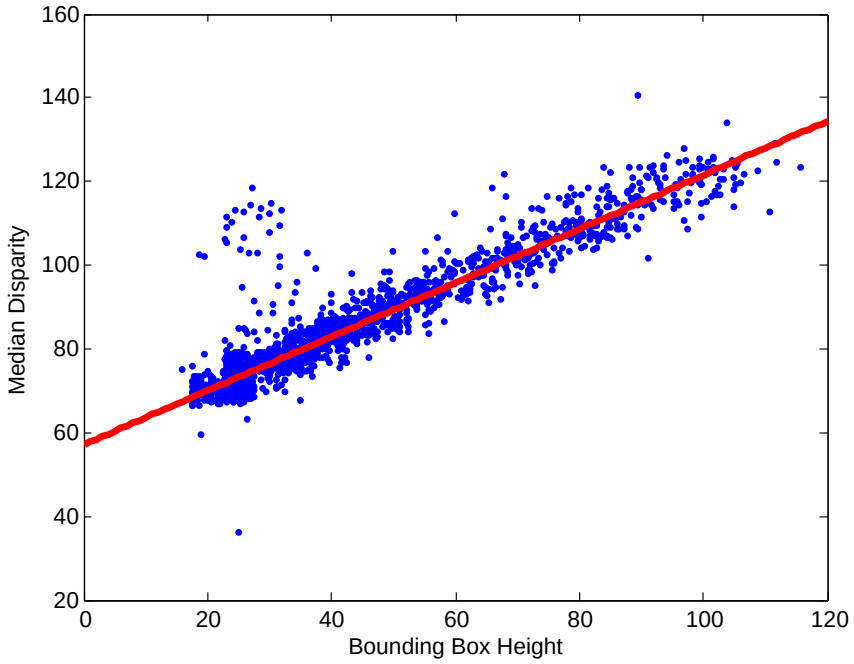


Figure VI.7: Linear relation of bounding box height and median disparity for positive samples of people. The data points are plotted in blue and the least-squares linear fit is plotted in red.

bounding box, we can then compute its distance from this ideal line as

$$\Delta L = \frac{|Ax + By + C|}{\sqrt{A^2 + B^2}} \tag{VI.1}$$

VI.D.4 Learning and Classification

While the single disparity feature could potentially be added to the color and infrared HOG features, it is likely that its power for classification would be lost in the hundreds of HOG features generated for each sample. Since ΔL arises from a different modality and attempts to model a completely different physical property, it is appropriate to treat these features independently. We build one classifier for the visual HOG features and another classifier for the disparity feature. Each classifier's result can than be probabilistically combined to determine the final classification.

We train a person classifier using the HOG features from the color and/or infrared imagery using a support vector machine (SVM) using radial basis function kernels [68]. We use cross-validation during training to give probability estimates that a bounding box contains a person, $p(\text{Person}|\text{HOG})$.

We model the disparity-based classification as being normally distributed around the distance ΔL from the line learned in Figure VI.7. We compute the probability that a region contains a person given the ΔL as

$$p(\text{Person}|\Delta L) = \text{erfc}(\frac{\Delta L}{\sqrt{2}\sigma}) \tag{VI.2}$$

where erfc is the complementary error function and σ is the standard deviation control parameter of the modeled Gaussian.

By making an independence assumption, we can construct the final classification probability as

$$p(\text{Person}) = p(\text{Person}|\text{HOG})p(\text{Person}|\Delta L) \tag{VI.3}$$

The additional benefit of using these features in two separate classifiers is that the relatively fast disparity classifier can be used to reduce the number of bounding

boxes that need to be evaluated for the slower HOG-based classifier. For example, there are potentially on the order of 10^6 evaluations. In practice, we have found that this can be reduced by two orders of magnitude to 10^4 by only considering bounding boxes with high probability from the disparity classifier.

VI.E Experimental Framework

VI.E.1 Experimental Testbed and Image Acquisition

We need to establish a framework for experimenting and analyzing person surveillance detection approaches that will facilitate a direct, frame-by-frame comparison of the various approaches that combine color and infrared stereo imagery. We designed a custom rig, shown in Figure VI.8, consisting of a matched color stereo pair and a matched infrared stereo pair. The two pairs share identical baselines and have been aligned in pitch, roll and yaw to maximize the similarities in field of view. Such a rig will allow us to compare Color Stereo, Infrared Stereo, Trifocal Color Stereo + Infrared (CSI), Trifocal Infrared Stereo + Color (ISC), and Tetravision approaches to person detection. A four-input video capture card is used to acquire the images and a time-stamping synchronization routine is used to best align the asynchronously captured video sequences.

Calibration data was obtained by illuminating a checkerboard pattern with high intensity halogen bulbs so the checks would be visible in both color and infrared imagery and standard calibration techniques could be applied to obtain the intrinsic and extrinsic parameters of the cameras. Color stereo and infrared stereo calibration was obtained from the matched calibration points using the Matlab Camera Calibration Toolbox [29]. The same point sets are also used to estimate the trifocal tensor for the trifocal CSI and ISC cases.

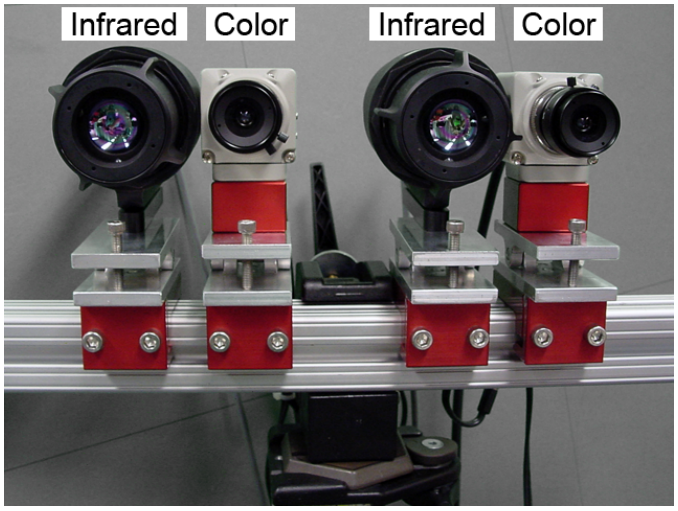


Figure VI.8: Experimental testbed: Two color cameras and two infrared cameras arranged in stereo pairs and mounted to the front of the LISA-P testbed.

VI.E.2 Data Set and Training

Videos were collected over several days in the scene shown in Figure VI.4. Twenty-one sequences of 352x240, 30 fps video were collected of different people moving throughout the environment at different times of day in an attempt to capture a wide range of illumination, position, occlusion and density conditions. Of those sequences, nineteen were used for annotation and training, while the remaining two were reserved as test sets. The two separate test videos were selected for their challenging and dense scenes and the fact that the people in the scene were not in the other videos. Cross-validation was not used in these experiments, as the resulting detections were evaluated by a human operator and increasing the number of test sequences would make the evaluation unmanageable. For each sequence, we compute color stereo, trifocal CSI, infrared stereo, and trifocal ISC variants of the original data using the dense disparity generation described in [63] with the trifocal tensor.

Annotation of Color Stereo and Trifocal CSI Data

When using the color stereo as the base reference image, we annotated 1654 positive samples of people in the scene. The positive samples range over 21 scales from 6-46 pixels wide. For each positive sample we attempt to obtain 10 negative samples by randomly translating the bounding box to a non-person region in the scene. We also obtain an additional 5 negative samples by randomly selecting a subregion of the positive bounding box as a negative sample. If a negative sample cannot be generated after a maximum number of iterations due to a dense scene, or if the subregion is smaller than the smallest person scale, we do not include that negative sample. In all, 22520 negative samples were gathered for training.

Annotation of Infrared Stereo and Trifocal ISC Data

We made every attempt to include identical samples for using the infrared stereo as reference as in the color stereo case. However, due to the slightly different fields of view, this was not always possible. Positive and negative samples were generated in the same manner as the color stereo case, resulting in 1425 positive and 19533 negative samples. We do maintain the same scale range of 6-46 pixels in bounding box width.

For training, the color and infrared parts of each sample are resized to 24×60 pixels. A $6 \times 15 \times 8$ dimensional HOG feature is computed for each of the color and infrared parts of the sample and used to train SVM classifiers with radial basis function (RBF) kernels. We use cross-validation of the training samples to obtain probability estimates for the classifiers. SVM classifiers are obtained for each of the four combinations of color and infrared imagery.

The training data is also used to learn the bounding box height-to-disparity function used to classify people in the disparity domain. We obtain a linear estimate of the function for both the color stereo and infrared stereo cases.

VI.F Experimental Evaluation

We analyze the reserved test sequences from the 21 training sequences. The sequences include various people moving through the scene with other moving objects including other vehicles and a dog. Detection was determined a success if the appropriately sized bounding box encapsulated the person in the scene. Naturally, false positives arose when bounding boxes did not encapsulate a person region and missed detection occurred when a person was not found by the classifier.

VI.F.1 Comparison

The sequences were evaluated for the color stereo, trifocal CSI, infrared stereo, and trifocal ISC. Additionally, we compare our trifocal approaches to the tetravision approach proposed by Bertozzi *et al.* [16]. The tetravision approach utilizes a four camera rig where color stereo and infrared stereo are analyzed independently and their detections combined to determine the overall detection. We apply this philosophy in our analysis by combining the results from our color stereo and infrared stereo using logical AND and OR operations on the bounding boxes.

Figure VI.9 shows the ROC curves for a sampled portion of the entire sequence. Plotted data points were generated by analyzing each classifier's detection/false positive rate when the detection probability threshold was set at 80, 85, 90, and 95 percent. Figure VI.10 shows example results of person detection using each of the compared approaches. In these examples, the detection probability was fixed at 90%.

Clearly, the two trifocal classifiers outperform the single modality classifiers by a large margin. For a false positive rate of 1 per frame, the multimodal classifiers increase the detection rate by over 45%, from 0.65 to almost 0.95. These are impressive gains, and while different feature selection and data profiles could yield more modest gains, there is clearly a substantial benefit in incorporating color and infrared features to create a superior discriminator of people in a scene. It is also clear that incorporating the color and infrared features for classification in this trifocal approach is better suited

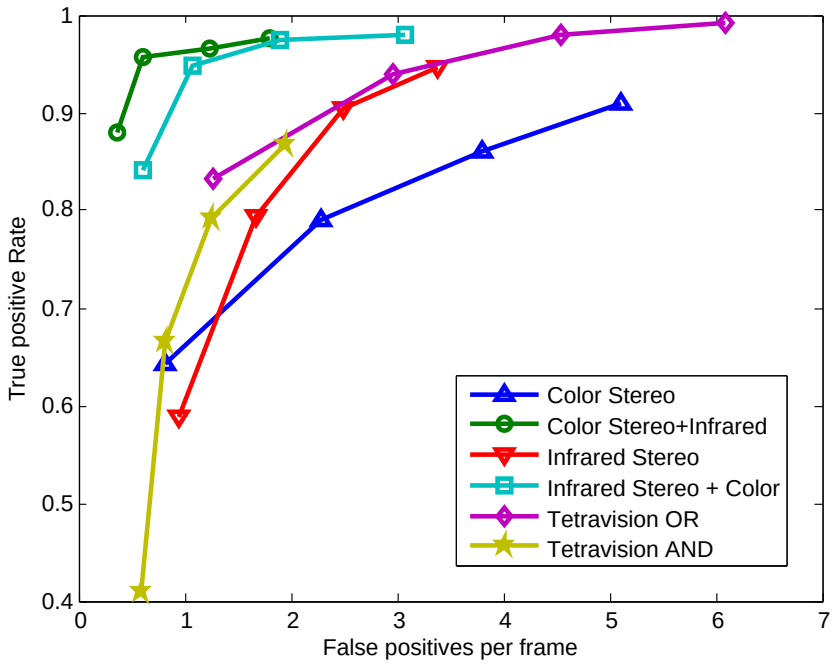


Figure VI.9: ROC curve of person detection using color/infrared SVM with disparity-based classifiers.

to detecting pedestrians than the independent classification and merge philosophy of the tetravision approach. Again, for a false positive rate of 1 per frame, we see an increase in detection of almost 20%. Incorporating the multispectral features at the classifier level will yield much more accurate detection than combining the detection results independently.

Of note is that color stereo is outperformed by infrared stereo, yet trifocal CSI performs better than trifocal ISC. While this may seem counterintuitive, a careful analysis can illuminate the cause for this swap in comparative performance. Since we use the stereo disparities to initially thin the person candidates in the scene, it is not surprising to see the infrared stereo outperform the color stereo case. The disparity generation algorithm we used [63] relies on windowed correlation matching, where highly textured areas are more easily matched than areas of low texture. Since infrared imagery is inherently low textured, the non-person regions produce fewer valid disparities, resulting in fewer false positives. Similarly, the areas that contain people are likely to have valid dis-

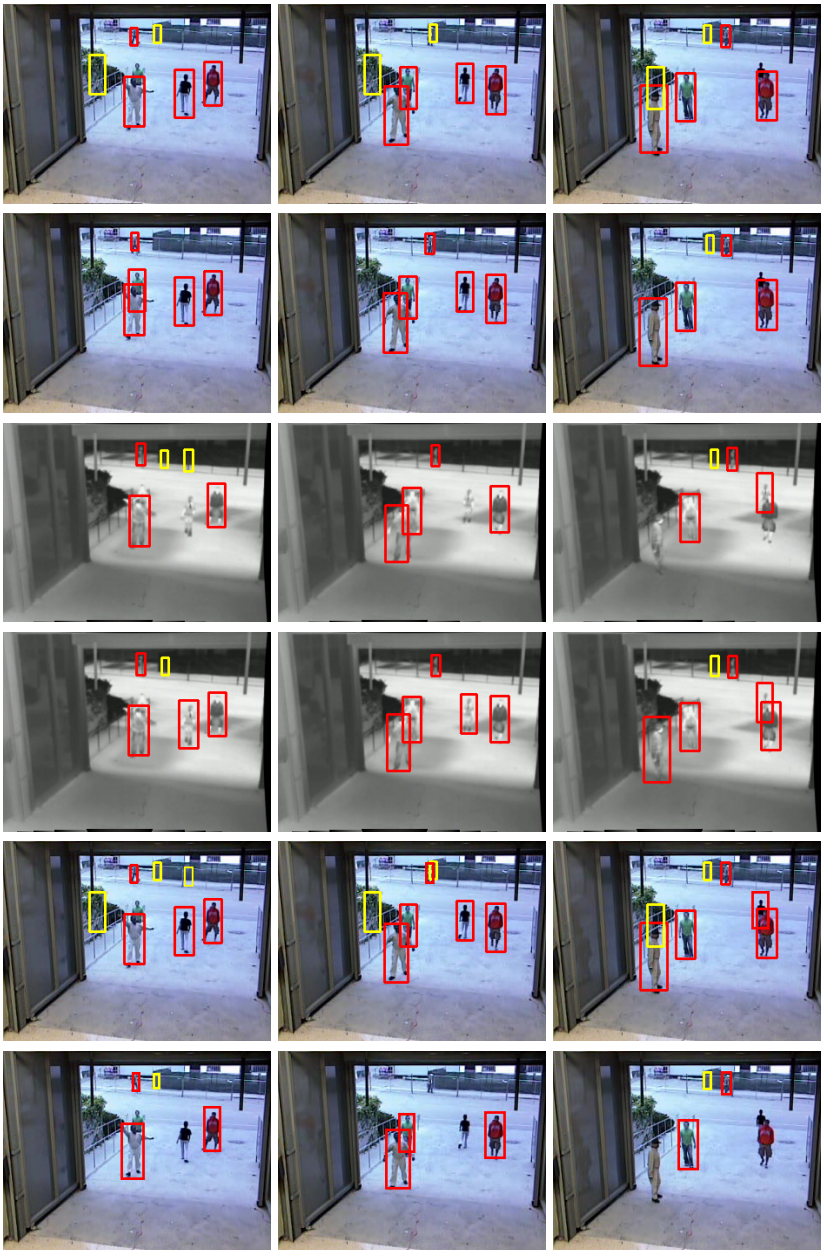


Figure VI.10: Example results of a frame-by-frame comparison of the person detection results using different combinations of color, infrared and disparity features. Successful detections are shown in red, false positives in yellow. Each row shows color stereo, trifocal CSI, infrared stereo, trifocal ISC, tetravision OR and tetravision AND.

parity estimates and will stand out in the imagery. The color imagery is often similarly textured in person or non-person regions, so more candidates are generated, increasing the potential for false positives.

However, the opposite is true when both the color and infrared is used in the SVM classifier. In this case, the increased disparity resolution and accuracy from the color stereo imagery allows for more accurate trifocal registration and improves the selection of candidate person regions. This improved fidelity makes it easier for the color and infrared trained SVM to differentiate person and non-person regions and yields the higher performance shown in the ROC curves. We also expect that when the trifocal registered modality is not available (e.g. color at night or infrared during poor temperature conditions), the detection rates will move towards their unimodal counterparts.

VI.F.2 Extended Analysis of Trifocal Detectors

We further focus our analysis on the top performing classifiers. Figure VI.11 shows successful detection results for example frames using the trifocal ISC framework. Figure VI.12 shows examples of successful detection for example frames using the best performing trifocal CSI framework. Notice how the framework yields accurate detection across a wide range of person scales, from people very large and near to the foreground to barely visible people deep in the background. These figures also demonstrate the classifiers' ability to suppress false positives from other objects in the scene, including vehicles and dogs.

Table VI.1 expands the analysis of the best performing trifocal CSI classifier by including several additional analyzed video sequences. The resulting analysis reinforces the results of the comparative analysis in Figure VI.9, showing an overall detection rate of 92.15% with 0.606 false positives per frame. This consistency further emphasizes the benefits of utilizing the trifocal CSI framework.

While the resulting detection rate is relatively high, we also achieve a seemingly high false positive rate of 0.606 false positives per frame (FPP). However, the SVM was trained to minimize the number of false postives per evaluated candidate win-

Figure VI.11: Good Results for Trifocal ISC

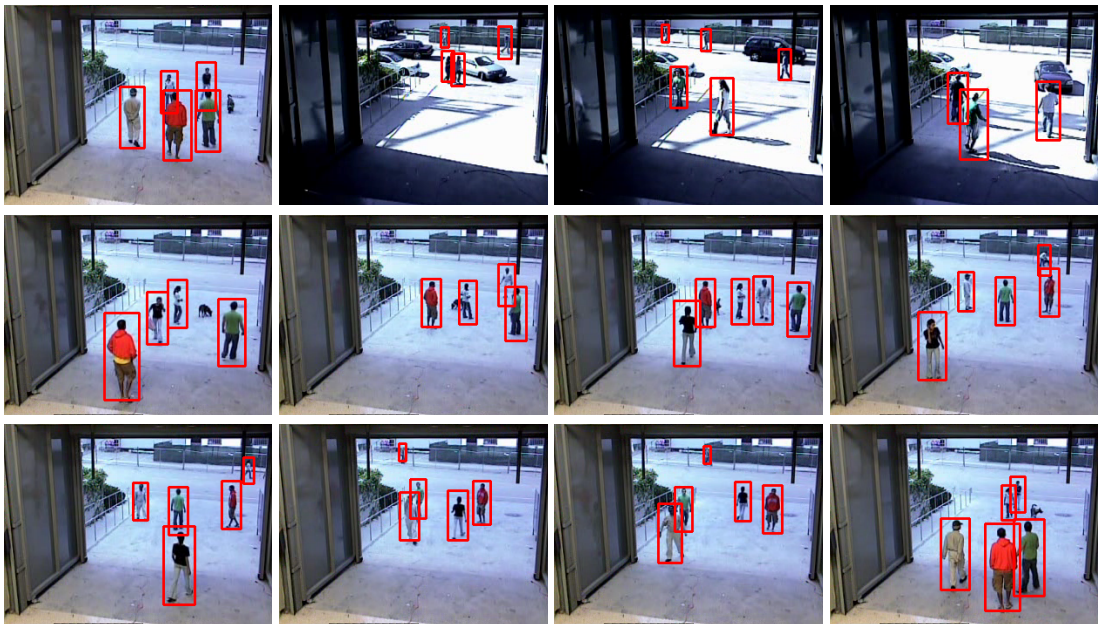


Figure VI.12: Good Results for Trifocal CSI

Table VI.1: Person Detection for Trifocal CSI Framework at 90% Threshold

	Video Sequence				
	1	2	3	4	Total
Total People	570	1240	810	425	3045
Total Frames	200	341	175	116	832
Detected People	535 93.86%	1130 91.13%	748 92.35%	393 92.47%	2806 92.15%
False Positives	103 0.515	209 0.613	151 0.863	43 0.371	506 0.606

Figure VI.13: Common false positive regions for trifocal CSI, shown in yellow.

dow sample (FPW). For each frame, we evaluate $352 \times 240 \times 21$ windows in the image, meaning our false positve rate per window (FPW) is 3.4×10^{-7}. Figure VI.13 shows examples of the false positives generated in our detection framework. The false positives in the images are shown in yellow. Our analysis has shown that an overwhelming majority of the generated false positives are located in the areas shown in these examples. A refinement to our approach could be to bootstrap these and other repeated false positives examples and retrain the SVM to achieve a lower false positive rate.

VI.F.3 Testing in different environments

Experiments were also conducted in another environment to test the trained person classifier's robustness to variations in scene perspective, background, density and lighting conditions. Data was collected in a new, outdoor environment that included multiple pathways through a grassy mall. Six video sequences were collected over several hours from two distinct perspectives that allowed for the capture of the natural

(a) Trifocal CSI (b) Trifocal ISC

Figure VI.14: Detection in crowded scene.

movement of people in the environment. In general, these sequences were denser in numbers of people than the initial experiments.

The sequences were evaluated for the best performing trifocal-based classifiers in the initial tests. The SVM classifiers used to evaluate the sequences were identical to those in the original tests. The disparity-based classifier was retrained to account for the change in disparity-to-bounding box size function in the new perspective. This can be done quickly by annotating a handful of new examples and estimating the new best linear fit.

Figure VI.14 shows an example of one of the densest frames in the test sequence, where 13 people occupy the scene. The trifocal CSI classifier is able to successfully detect every person with no false positives, while the trifocal ISC classifier detects all but a single pedestrian, again with no false positive. We emphasize that no additional samples were used to evaluate these sequences and many of the objects in the scene, such as the grass, tree and foreground fencing have not been modeled explicitly by the SVM classifiers. Figure VI.15 and Figure VI.16 show additional detection examples for the trifocal ISC and CSI cases, respectively.

We compiled a comparison of the trifocal detection results for a test sequence in Table VI.2. The results are on par with the original series of test sequences. We

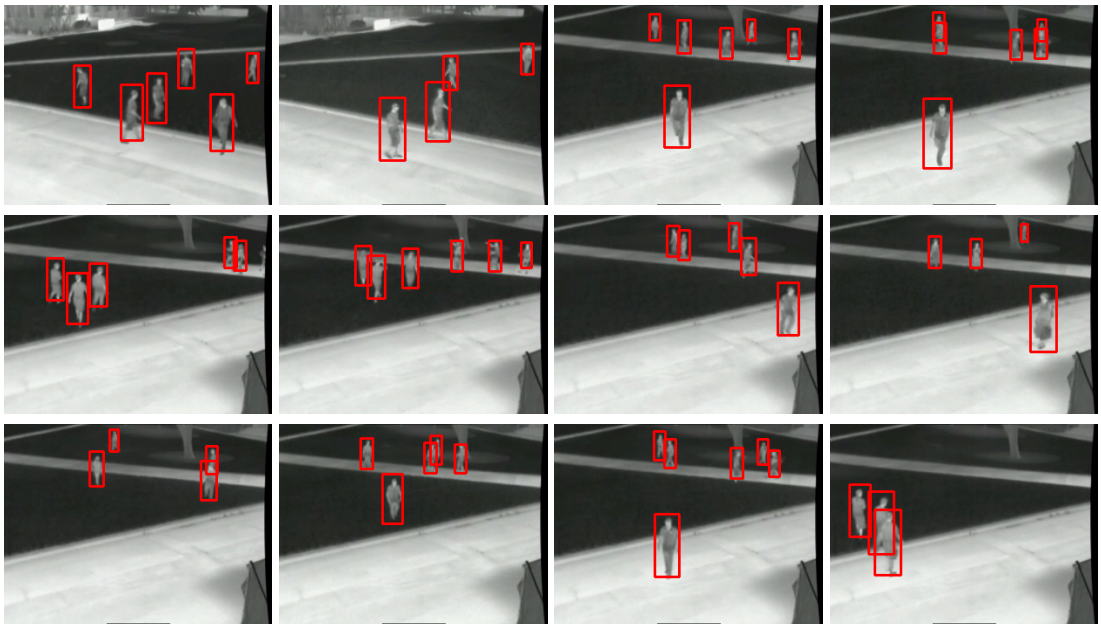


Figure VI.15: Additional Results for Trifocal ISC

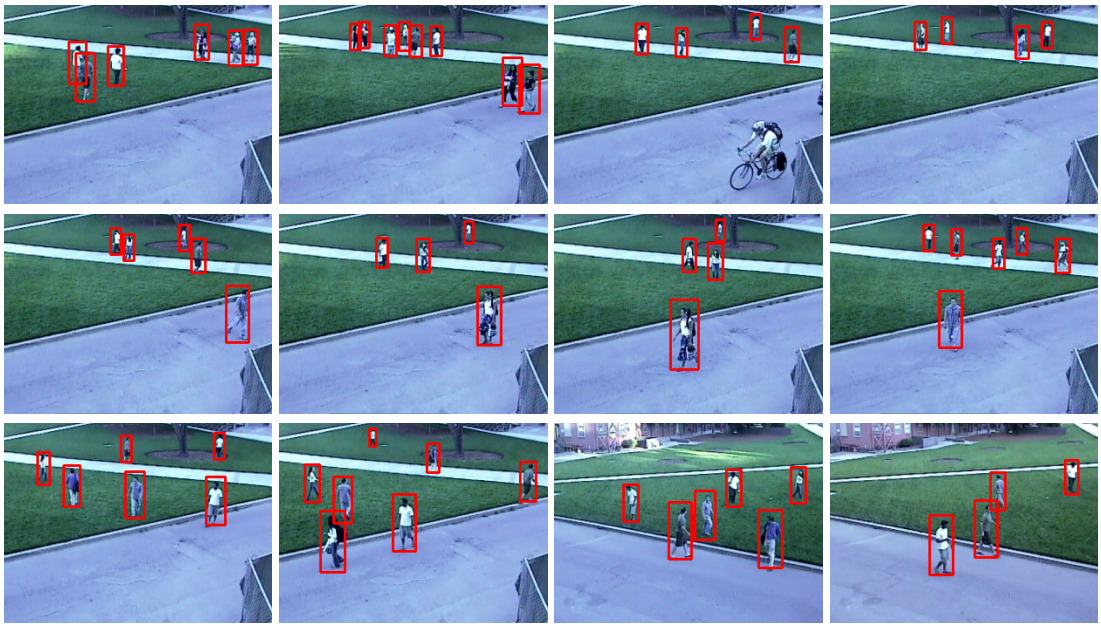


Figure VI.16: Additional Results for Trifocal CSI

Table VI.2: Person Detection Comparison for Trifocal CSI and ISC at 90% Threshold

	Classifier Type	
	Trifocal CSI	Trifocal ISC
Total People	1019	995
Total Frames	130	130
Detected People	876 86.00%	798 80.20%
False Positives	107 0.860	90 0.692

do see a noticeable decrease in the per frame detection rate. This is likely due to the incorporation of a new scene that has no support in the trained classifier. Additionally, these new test sequences have, on average, twice as many people in the scene as the original test sequences. This increases the occurrence of occlusion that can lead to a missed detection of a person for an individual frame.

VI.F.4 Temporal Filtered Detection and Tracking

We believe that these per frame detection rates we achieve are really the lower bound, and that increased performance can come from the temporal analysis of the per frame detections. In our analysis, we consider a missed detection for any frame where a person was not properly encapsulated by a bounding box. However a single missed detection for a person in a given frame is usually corrected in the next few frames. Such a missed detection can be thought of as a missing data sample in a larger tracking framework.

Figure VI.17 shows a timelapsed image of a typical 60 frame sequence in our experiments, where the start and end frames overlayed on each other. For each person in the scene, we plot the correct per frame detections in solid dots (blue, cyan, red, magenta and green, respectively) and plot the missed detections in yellow circles. This plot demonstrates how the intermittent missed detections would not detract from an overall

Figure VI.17: Time lapse display of typical experimental sequence with per frame detection overlayed. Correct per frame detections are shown in colored dots and missed detections are indicated as yellow circles.

tracking framework. The solid dots for each person clearly indicate the path taken by each person and the missed detections are relatively few. This means that the missed detection rate is mostly due to intermittent missing data points for tracking rather than being completely unable to detect a person in the scene. Temporal analysis is a crucial aspect of algorithmic approaches to surveillance, as the movement and interaction of objects in the scene can give fundamental insight to the situational analysis of the scene [65]. We feel that our trifocal classification approach gives a natural and robust input to common person tracking techniques such as Kalman [75] and Particle Filtering [76].

VI.G Summary

We have presented a methodology for analyzing multimodal and multiperspective systems for person surveillance. By incorporating an experimental testbed consisting of two color and two infrared cameras, we are able to expand multispectral color and infrared analysis beyond the specialized long-range surveillance experiments of previous approaches to more general scene configurations common to unimodal approaches.

We presented an algorithmic framework for detecting people in a scene that probabilistically combines a support vector machine trained on histogram of oriented gradient (HOG) features extracted from color and infrared images with a detector based on the relationship between person size and depth in the scene to create a disparity-based detector. This framework was used to train person detectors for the various combinations of color and infrared multiperspective imagery, including color stereo, infrared stereo, tetravision and trifocal tensor configurations.

The trained detectors could then be used in an experimental evaluation of video sequences captured with our designed test bed. The evaluation definitively demonstrates the performance gains achievable when using the trifocal framework to combine color and infrared features in a unified framework. Both of the trifocal setups outperform their unimodal equivalents, as well as the tetravision based analysis. Our experiments

also demonstrate how the trained detector generalizes well to different scenes and can provide robust input to an additional tracking framework.

The text of this chapter, in part, is a reprint of the material as it appears in: Stephen J. Krotosky and Mohan M. Trivedi, "Algorithmic Framework and Experimental Evaluation for using Multiperspective Color and Infrared Features for Person Surveillance", *IEEE Trans. on Circuits and Systems for Video Technology*, submitted. I was the primary researcher of the cited material and the co-author listed in this publication directed and supervised the research which forms a basis for this chapter.

Chapter VII

Conclusions

Detecting and tracking people has attracted a lot of research interest in the computer vision community. Traditionally, efforts had been made to do detection using only monocular color 2D imagery. Efforts were pushed and eventually 3D and stereo approaches were incorporated to help detect and track through difficult scenarios such as occlusion and large scale variations. As thermal infrared imagery has become more viable and affordable, algorithms have been developed to exploit its properties for detecting people in both monocular and stereo imagery. More recent efforts have made attempts to combine color and infrared analysis, but have been limited in both experimental scope and scene generality. This dissertation has been devoted to expanding these to achieve a general framework for multiperspective analysis of color and infrared imagery by developing algorithms to 1) solve the problem of registering color and infrared imagery and 2) create a unified framework for detecting people in color and infrared multiperspective imagery.

We presented the related studies of multimodal image registration and categorized the registration methodologies into four distinct sectors based on the assumptions about scene configuration in order to examine how the current state of literature fails to accommodate registration for general scene. We also examined state-of-the-art stereo algorithms that are designed to handle correspondence matching for unmatched image data and definitively show that these approaches are unsuitable for finding correspondence in cross-spectral stereo imagery, where a color and infrared camera are joined in a stereo pair. As an alternative, we propose a region-based approach to correspondence matching that is able to successfully perform correspondence matching by relying on an initial segmentation and disparity voting-based methodology to registering foreground objects in the scene. This, we believe, is the first such algorithm capable of registering color and infrared imagery using only two cameras. While our algorithm does not completely achieve the desired generality, we feel we have laid the groundwork for future exploration and development by giving an extensive review of the issues and challenges of registering the imagery and elucidating the desirable properties of features that could help improve the generality of our approach.

Extensive experimental evaluations of our proposed cross-spectral stereo registration algorithm were performed. We presented experimental studies in registering people in both indoor surveillance from a static camera and outdoor pedestrian detection from a moving vehicle. We also offer a comparison of our approach to ground truth and the current state of related studies, with both ideal and realistic initial segmentations. We also experimentally validate the robustness of our approach by evaluating additional data taken from different cameras in another environment. Finally, we show how our approach to cross-spectral stereo registration can be used to track people in a 3D context. The experimental results demonstrated the ability of our algorithm to register foreground objects in the scene and achieved a level of registration accuracy and robustness better than current state-of-the-art approaches.

Our study then focused on studying how color and infrared imagery can be used to improve person detection algorithms. In the context of pedestrian detection, we first compared how the disparity information from color stereo and infrared stereo can be used to detect potential objects in the scene. The high success of the disparity information from both modalities motivated a discussion of the color and infrared features that can be extracted to further classify the potential objects into pedestrian and non-pedestrian regions. This lead to the development of our experimental framework that allows us to compare pedestrian classifiers utilizing all combinations of color, infrared and disparity features. We also propose a trifocal framework consisting of a color stereo camera rig combined with an infrared camera in order to quickly register the multimodal data for our analysis.

We extend the analysis of multispectral and multiperspective approaches to person detection in the context of surveillance. We further justify our trifocal approach to registration by demonstrating its superiority to the planar homography approach in terms of scene generality and robustness. The trifocal approach is able to register any object in the scene that is able to be registered in stereo imagery. This allows general scene configurations and also allows for a direct comparison to conventional monocular and unimodal stereo approaches. With this in mind, we present a framework for person

detection that can combine color, infrared and disparity features in a unified manner and expands the robustness and accuracy of the method proposed in the previous chapter. We then use this algorithmic framework to present a detailed comparison of person detection using various combinations of color, infrared and disparity features.

We have definitively demonstrated that our unified trifocal framework easily outperforms both unimodal stereo analysis and multimodal "tetravision" analysis that separately combines color and infrared stereo analysis and that such an approach is currently the most effective and efficient way to achieve color and infrared analysis in a general scene configuration. We present extensive evaluation of the trifocal-based experiments to illustrate the improved detection rates that can be achieved when incorporating multispectral data in the detection framework. The trifocal framework we developed could potentially be deployed in many person detection application realms to improve the overall detection performance of current approaches including person surveillance, pedestrian detection and intelligent environments.

Bibliography

[1] P. Thevenaz, M. Bierlaire, and M. Unser, "Halton sampling for image registration based on mutual information," *Sampling Theory in Signal and Image Processing*, In Press.

[2] M. M. Trivedi, S. Y. Cheng, E. M. C. Childers, and S. J. Krotosky, "Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation," *IEEE Trans. Veh. Technol.*, vol. 53, no. 6, pp. 1968–1712, Nov. 2004.

[3] J. Davis and V. Sharma, "Fusion-based background-subtraction using contour saliency," in *IEEE CVPR Workshop on Object Tracking and Classification beyond the Visible Spectrum*, 2005.

[4] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[5] G. Egnal, "Mutual information as a stereo correspondence measure," University of Pennsylvania, Tech. Rep. MS-CIS-00-20, 2000.

[6] P. Thevenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. Image Processing*, vol. 9, no. 12, pp. 2083–9, Dec. 2000.

[7] G. L. Foresti, C. S. Regazzoni, and P. K. Varshney, *Multisensor Surveillance Systems: The Fusion Perspective*. Springer Press, 2003.

[8] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2002.

[9] C. O'Conaire, E. Cooke, N. O'Connor, N. Murphy, and A. Smeaton, "Background modeling in infrared and visible spectrum video for people tracking," in *IEEE CVPR Workshop on Object Tracking and Classification beyond the Visible Spectrum*, 2005.

[10] M. Irani and P. Anandan, "Robust multi-sensor image alignment," in *Computer Vision, 1998. Sixth International Conference on*, 1998.

[11] E. Coiras, J. Santamaria, and C. Miravet, "Segment-based registration technique for visual-infrared images," *Optical Engineering*, vol. 39, no. 1, pp. 282–289, Jan. 2000.

[12] J. Han and B. Bhanu, "Detecting moving humans using color and infrared video," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2003.

[13] M. Itoh, M. Ozeki, Y. Nakamura, and Y. Ohta, "Simple and robust tracking of hands and objects for video-based multimedia production," in *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2003.

[14] G. Ye. (2005) Image registration and super-resolution mosaicing. http://www.library.unsw.edu.au/∼thesis/adt-ADFA/uploads/approved/adt-ADFA20051007.144609/public/01front.pdf.

[15] X. Ju, J.-C. Nebel, and J. P. Siebert, "3D thermography imaging standardization technique for inflammation diagnosis," in *Proceedings of SPIE, Photonics Asia*, 2004.

[16] M. Bertozzi, A. Broggi, M. Felias, G. Vezzoni, and M. D. Rose, "Low-level pedestrian detection by means of visible and far infra-red tetra-vision," in *IEEE Conference on Intelligent Vehicles*, 2006.

[17] H. Chen, P. Varshney, and M. Slamani, "On registration of regions of interest (ROI) in video sequences," in *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03)*, 2003.

[18] J. Davis and V. Sharma, "Robust detection of people in thermal imagery," in *IEEE 17th International Conference on Pattern Recognition*, 2004.

[19] ——, "Robust background-subtraction for person detection in thermal imagery," in *Computer Vision and Pattern Recognition Workshop, 2004 Conference on*, 2004.

[20] S. J. Krotosky and M. M. Trivedi, "Mutual information based registration of multimodal stereo videos for person tracking," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, May-Jun. 2007.

[21] J. Kim, V. Kolmogorov, and R. Zabih, "Visual correspondence using energy minimization and mutual information," in *Ninth IEEE International Conference on Computer Vision*, 2003.

[22] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Computer Vision and Pattern Recognition*, 2005.

[23] D. Scharstein and R. Szeliski. (2005) Middlebury College stereo vision research page. [Online]. Available: http://cat.middlebury.edu/stereo/

[24] S. Marapane and M. M. Trivedi, "Multi-primitive hierarchical (MPH) stereo analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, no. 3, pp. 227–240, Mar. 1994.

[25] S. Marapane and M. Trivedi, "Region-based stereo analysis for robotic applications," *IEEE Trans. Syst., Man, Cybern., Special Issue on Computer Vision*, vol. 19, no. 6, pp. 1447–1464, 1989.

[26] L. Cohen, L. Vinet, P. Sander, and A. Gagalowicz, "Hierarchical region based stereo matching," in *Computer Vision and Pattern Recognition*, 1989.

[27] Y. Wei and L. Quan, "Region-based progressive stereo matching," in *Computer Vision and Pattern Recognition*, 2004.

[28] M. Bleyer and M. Gelautz, "Graph-based surface reconstruction from stereo pairs using image segmentation," *Proc. SPIE*, vol. 5665, pp. 288–299, Jan. 2005.

[29] J.-Y. Bouguet. Camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/.

[30] S. J. Krotosky and M. M. Trivedi, "Registration of multimodal stereo images using disparity voting from correspondence windows," in *IEEE Conference on Advanced Video and Signal based Surveillance (AVSS'06)*, 2006.

[31] ——, "Multimodal stereo image registration for predestrian detection," in *IEEE Conference on Intelligent Transportation Systems*, 2006.

[32] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imaging*, vol. 11, no. 3, pp. 163–256, June 2005.

[33] A. S. Ogale and Y. Aloimonos, "A roadmap to the integration of early visual modules," *International Journal of Computer Vision: Special Issue on Early Cognitive Vision*, 2006.

[34] T. B. Moesland and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[35] M. Harville and D. Li, "Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[36] K. Huang and M. M. Trivedi, "Video arrays for real-time tracking of person, head, and face in an intelligent room," *Machine Vision and Applications*, vol. 14, no. 2, pp. 103–111, June 2003.

[37] S. J. Krotosky and M. M. Trivedi, "On color, infrared and multimodal stereo approaches to pedestrian detection," *IEEE Trans. Intell. Transport. Syst.*, IN PRESS.

[38] http://www.worldbank.org/html/fpd/transport/roads/safety.htm.

[39] Traffic safety facts 2004: A compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system. National Highway Traffic Safety Association, US Dept. of Transportation. [Online]. Available: http://www-nrd.nhtsa.dot.gov/pdf/nrd-30/NCSA/TSFAnn/TSF2004.pdf

[40] J. R. Crandall, K. S. Bhalla, and N. J. Madeley, "Designing road vehicles for pedestrian protection," *British Medical Journal*, vol. 324, no. 7346, pp. 1145–1148, May 2002.

[41] S. K. Singh, "Review of urban transportation in India," *Journal of Public Transportation*, vol. 8, no. 1, pp. 79–97, 2005.

[42] D. Mohan, "Traffic safety and health in Indian cities," *Journal of Transport and Infrastructure*, vol. 9, no. 1, pp. 79–94, 2002.

[43] Y. Fang, K. Yamada, Y. Ninomiya, B. Horn, and I. Masaki, "Comparison between infrared-image-based and visible-image-based approaches for pedestrian detection," in *IEEE Intelligent Vehicles Symposium*, 2003.

[44] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intell. Transport. Syst.*, 2007.

[45] L. Andreone, F. Bellotti, A. D. Gloria, and R. Laulette, "SVM-based pedestrian recognition on near-infrared images," in *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, 2005.

[46] H. Cheng, N. Zheng, and J. Qini, "Pedestrian detection using sparse gabor filters and support vector machine," in *IEEE Conference on Intelligent Vehicles*, 2005.

[47] A. Shashua, Y. Gdalyahu, and G. Hayun, "Pedestrian detection for driving assistance systems: Single-frame classification and system level performance," in *IEEE Conference on Intelligent Vehicles*, 2004.

[48] Y. Wi, T. Yu, and G. Hua, "A statistical field model for pedestrian detection," in *Computer Vision and Pattern Recognition*, 2005.

[49] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Computer Vision and Pattern Recognition*, 2005.

[50] S. Munder and D. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 11, pp. 1863–1868, Nov. 2006.

[51] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Trans. Intell. Transport. Syst.*, vol. 6, no. 1, pp. 63–71, Mar. 2005.

[52] A. Broggi, A. Fascioli, P. Grisleri, T. Graf, and M. Meinecke, "Model-based validation approaches and matching techniques for automotive vision based pedestrian detection," in *Computer Vision and Pattern Recognition*, 2005.

[53] Y. Fang, K. Yamada, Y. Ninomiya, B. K. P. Horn, and I. Masaki, "A shape-independent method for pedestrian detection with far-infrared images," *IEEE Trans. Veh. Technol.*, vol. 53, no. 6, pp. 1679–1697, Nov. 2004.

[54] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," in *IEEE Conference on Intelligent Vehicles*, 2006.

[55] S. Cheng and M. M. Trivedi, "Turn-intent analysis using body pose for intelligent driver assistance," *IEEE Pervasive Computing*, vol. 5, no. 4, pp. 28–37, Oct.-Dec. 2006.

[56] M. Bertozzi, A. Broggi, C. Caraffi, M. D. Rose, M. Felisa, and G. Vezzoni, "Pedestrian detection by means of far-infrared stereo vision," *Computer Vision and Image Understanding*, vol. 106, no. 2, 2007.

[57] M. Szarvas, A. Yoshizawa, M. Yamamoto, , and J. Ogata, "Pedestrian detection with convolutional neural networks," in *IEEE Intelligent Vehicles Symposium*, 2005.

[58] L. Zhao and C. Thorpe, "Stereo and neural network-based pedestrian detection," *IEEE Trans. Intell. Transport. Syst.*, vol. 1, no. 3, pp. 148–154, Sept. 2000.

[59] G. Grubb, A. Zelinsky, L. Nilsson, and M. Rilbe, "3D vision sensing for improved pedestrian safety," in *IEEE Conference on Intelligent Vehicles*, 2004.

[60] M. A. Sotelo, I. Parra, D. Fernandez, and E. Naranjo, "Pedestrian detection using svm and multi-feature combination," in *IEEE Conference on Intelligent Transportation Systems*, 2006.

[61] X. Lie and K. Fujimura, "Pedestrian detection using stereo night vision," *IEEE Trans. Veh. Technol.*, vol. 53, no. 6, pp. 1657–1665, Nov. 2004.

[62] R. Labayrade, D. Aubert, and J.-P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation." in *IEEE Conference on Intelligent Vehicles*, 2002.

[63] K. Konolige, "Small vision systems: hardware and implementation," in *Eighth International Symposium on Robotics Research*, 1997.

[64] M. M. Trivedi, T. Gandhi, and J. McCall, "Looking-in and looking-out of a vehicle: Computer-vision-based enhanced vehicle safety," *IEEE Trans. Intell. Transport. Syst.*, Mar. 2007.

[65] S. Park and M. M. Trivedi, "Multi-person interaction and activity analysis: A synergistic track- and body- level analysis framework," *Machine Vision and Applications: Special Issue on Novel Concepts and Challenges for the Generation of Visual Surveillance Systems*, 2007.

[66] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, 2005.

[67] L. Zhang, B. Wu, and R. Nevatia, "Pedestrian detection in infrared images based on local shape features," in *Computer Vision and Pattern Recognition*, 2007.

[68] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[69] T. Gandhi and M. M. Trivedi, "Vehicle surround capture: Survey of techniques and a novel omni-video-based approach for dynamic panoramic surround maps," *IEEE Trans. Intell. Transport. Syst.*, vol. 7, no. 3, pp. 293–308, Sept. 2006.

[70] J. McCall, D. Wipf, M. Trivedi, and B. Rao, "Lane change intent analysis using robust operators and sparse bayesian learning," *IEEE Trans. Intell. Transport. Syst.*, 2007.

[71] J. McCall and M. Trivedi, "Driver behavior and situation aware brake assistance for intelligent vehicles," *Proceedings of IEEE, Special Issue on Advanced Automobile Technologies*, Feb. 2007.

[72] Y. Ran, I. Weiss, Q. Zheng, and L. Davis, "Pedestrian detection via periodic motion analysis," *International Journal of Computer Vision*, vol. 71, no. 2, pp. 143–160, 2007.

[73] J. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, May 2007.

[74] A. Leykin, Y. Ran, and R. Hammoud, "Thermal-visible video fusion for moving target tracking and pedestrian classification," in *Computer Vision and Pattern Recognition*, 2007.

[75] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME - Journal of Basic Engineering*, vol. 10, pp. 35–45, 1960.

[76] A. Doucet, C. Andrieu, and S. Godsill, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.