2009-03-20

Combined Visible and Infrared Video for Use in Wilderness Search and Rescue

Nathan D. Rasmussen
Brigham Young University - Provo

COMBINED VISIBLE AND INFRARED VIDEO FOR USE IN WILDERNESS SEARCH AND RESCUE

by

Nathan D. Rasmussen

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computer Science

Brigham Young University

August 2009

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Nathan D. Rasmussen

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Bryan S. Morse, Chair
Date	Michael A. Goodrich
Date	Kent E. Seamons

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Nathan D. Rasmussen in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date	Bryan S. Morse Chair, Graduate Committee

Accepted for the Department

Date	Kent E. Seamons Graduate Coordinator

Accepted for the College

Date	Thomas W. Sederberg Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

COMBINED VISIBLE AND INFRARED VIDEO FOR USE IN WILDERNESS SEARCH AND RESCUE

Nathan D. Rasmussen

Department of Computer Science

Master of Science

Mini Unmanned Aerial Vehicles (mUAVs) have the potential to be a great asset to Wilderness Search and Rescue groups by providing a bird's eye view of the search area. These vehicles can carry a variety of sensors to better understand the world below. This paper proposes using both Infrared (IR) and Visible Spectrum cameras on a mUAV for Wilderness Search and Rescue. It details a method for combining the color and heat information from these two cameras into a single fused display to reduce needed screen space for remote field use. To align the video frames for fusion, a method for simultaneously pre-calibrating the intrinsic and extrinsic parameters of the cameras and their mount using a single multi-spectral calibration rig is also presented. A user study conducted to validate the proposed image fusion methods showed no reduction in performance when detecting objects of interest in the single-screen fused display compared to a side-by-side display. Furthermore, the users'

increased performance on a simultaneous auditory task showed that their cognitive load was reduced when using the fused display.

ACKNOWLEDGMENTS

I would like to give thanks to those who have helped me accomplish this work. Special thanks go to my advisor Dr. Bryan Morse who has assisted me and been patient as I have tried to push and get things done as quickly as possible. He has been a resource to help me with ideas and let me pursue various different ideas that have shaped this work. I thank Dr. Michael Goodrich for his assistance with the WiSAR research group. I would like to thank the many researchers previously and currently involved with the BYU MAGICC Lab. They have provided much expertise with the UAV's which have enabled this work to proceed. I would like to thank the Utah County Search and Rescue for their support in this research, providing us information and resources so that we can know what technologies could help in searches. I give special thanks to Ron Zeeman of this group for his continual support of our research group and his collaboration for finding out ways that we can work with Infrared cameras. I thank Daniel Thornton, Carson Fenimore, Brian Buss, Mike Roscheck, Doug Kennard, Brian Price, Stephen Cluff, Damon Gerhardt and the many other researchers in the Computer Graphics and Vision Lab as well as the Human-Centered Machine Intelligence Lab for their knowledge and assistance that helped shape my work. I would like to thank my friends and family for their continual encouragement that has helped me to continue on through the hard times.

I need to give special thanks to my wife Samantha. She has given me support throughout this entire process and helped me have extra motivation to get things

done. I am grateful for her willingness to go out and help me obtain the data I needed to accomplish this work.

I would like to thank my Heavenly Father for the inspiration I have felt that has helped me to overcome the various difficult problems I have encountered during this work. He has supported me and helped to open my mind to possible solutions when things were not going well.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Western United States is known by outdoor enthusiasts for its wilderness areas, but these areas do not exist without an element of danger. Too often people find themselves lost in wilderness areas and local Search and Rescue teams are called out to search for them. Time is critical for these individuals, as their chances of survival decrease the longer they are out in the elements. Searchers spend countless hours each year trying to find and help lost individuals. Search and Rescue teams utilize many specialists to assist in different types of searches, such as pilots who assist them from airplanes or helicopters. These aerial searchers speed up the search by being able to cover large areas in a short period of time. They also give a different perspective to the search given their bird's eye view, as they can see areas that could easily be missed by a searcher walking just feet away due to terrain, trees, or other obstructions.

Aerial searching comes, however, with a number of disadvantages. Conventional aircraft are very costly to purchase and operate, and pose potential danger to the pilot and crew. In the event of an accident, the Search and Rescue team now has two incidents they need to assist with rather than just one. In 2006 there was a case in Eastern Utah where a sheriff's deputy was killed during a search when the helicopter he was in hit power lines and crashed [1].

In recent years, there have been great advances in research surrounding the use of Unmanned Aerial Vehicles (UAVs) to obtain video from the air. This aerial

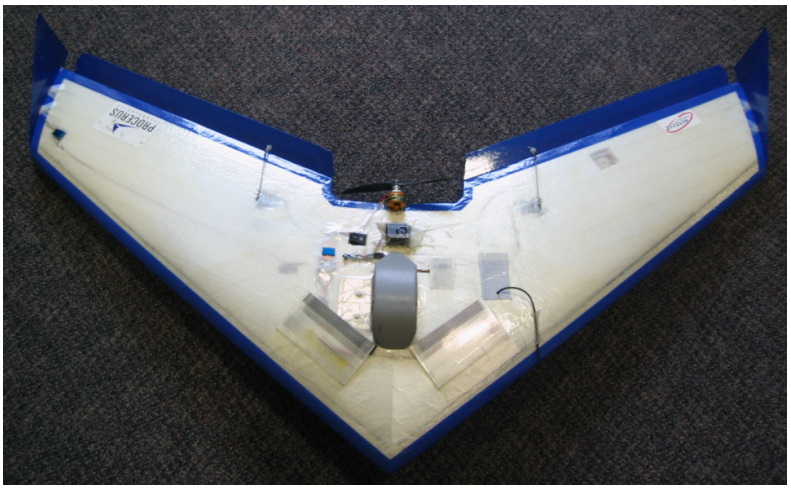


Figure 1.1: Mini Unmanned Aerial Vehicle

video can be used to replace aerial searchers and reduce the costs and dangers of searching from the air. Mini Unmanned Aerial Vehicles (mUAVs) have also been used because they bring additional advantages. mUAVs can be launched on location in many terrains. mUAVs cost far less to operate, making them more attainable for groups that have limited funding. mUAVs also reduce danger by being smaller and thus reducing the size of any affected area in the event there is a crash. Throughout the remainder of the paper, when we use the term UAV we are specifically referring to mini UAVs.

UAVs are capable of carrying a variety of sensors. The most common sensor carried by these is a visible-spectrum color camera, which we will refer to as a "Visible camera". These cameras provide views similar to those that pilots can obtain in manned aircraft (a sample frame is shown in Figure 1.2).

Another sensor that has potential to be very helpful in Wilderness Search and Rescue is an infrared (IR) camera (a sample frame is shown in Figure 1.3). This type of camera increases the information that an aerial searcher can obtain by being able to see the heat that our bodies produce. IR cameras are most useful during times or in areas when emitted body heat is greater than heat emitted from the surroundings, such as during the nighttime, in the early morning hours, over snow covered ground,

Figure 1.2: Image from a Visible Spectrum Camera

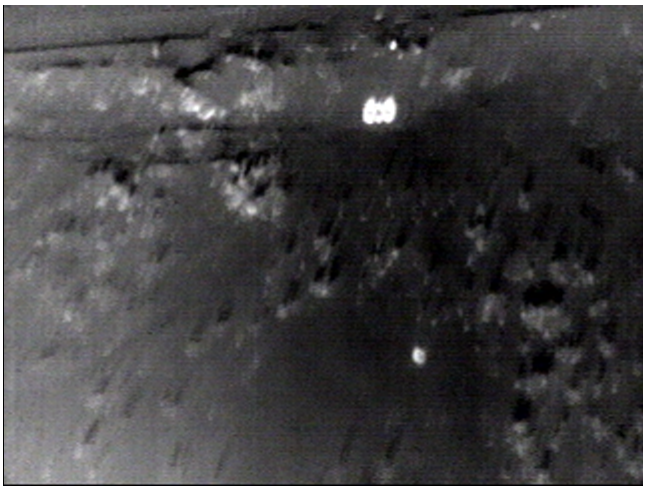


Figure 1.3: Image from a Infrared Spectrum Camera.

or over water. In many of these conditions, IR cameras could be used in conjunction with Visible cameras to maximize the information that the searchers receive.

Video coming from UAVs does not come without drawbacks. The aerial video gives a lot of information to the users all at once. Each frame has the potential of containing new information that needs to be analyzed, yet the viewer only has a fraction a second to do so. The amount of information is greatly increased when multiple video sensors are used. Significant screen real-state is needed to display all

of this information, making a small portable system for use in Wilderness Search and Rescue difficult to build or use.

UAVs equipped with Visible and IR cameras have great potential to be an asset to Search and Rescue teams. We have developed a method for solving the problems of screen real-estate and information overload introduced by working with two cameras. This is done by combining the information from the two cameras into a single view, retaining enough information from both image modalities that searchers are able to successfully find people and objects of interest. To be capable of combining the information we have also developed a method for calibrating the internal and external parameters of our multi-modal cameras and their mount so that we can align the video frames coming from the two cameras.

1.1 Related Work

The area of search and surveillance using UAVs has seen an expansion in recent years due to the increasing availability of UAVs (for example [2, 3, 4]). The research done in this area has covered a variety of topics, many of which are aimed at helping people who are working with the video.

1.1.1 Enhancement of Single Optical Video Streams

One basic task that can be performed to help the users better understand the video is to stabilize it [5, 6, 7]. Video stabilization decreases the jitter found in the videos and reduces the problems users have watching the video for long segments of time.

To help searchers have better awareness of the surroundings, other researchers have used global mosaics [5, 8, 9]. Mosaics are built by finding the alignment from one frame to the next and then warping all of the video's frames (using these alignments) together onto a large canvas. These warps accumulate propagated error that needs to be corrected. This error is taken out by doing an error minimization on the

entire system correcting each individual alignment warp. This large scale error minimization, called a bundle adjustment, cannot be performed in real-time on consumer hardware.

To reduce the need for full-scale error minimization, local mosaics can be used to aid in local understanding. Morse et al. [10] compared stabilization to using local mosaics to determine which would provide increased target detection rates. They found that local mosaics provided a larger temporal window that removed the need to see the target immediately allowing the user to look in the mosaic history and see objects.

The alignment of video frames used to create mosaics leads to another area in which researchers have worked with super resolution [2, 11] to assist users. Super resolution is a technique where the system exploits the small inconsistencies in what is seen to create an image that has a higher resolution than the original by using the small changes in pixel values to fill in the information gaps in the lower-resolution images. This method works well for taking a close look at a single item to get more detail and information but does not simplify the display for the user.

While each of these techniques has the potential of helping with a single video stream, none of them naturally extend to assisting the user with multiple streams.

1.1.2 Vision techniques using IR cameras

Infrared video images a different area of the electro-magnetic spectrum and has a number of different uses. Researchers have also used UAVs equipped with infrared for detecting fire [12, 13, 14]. These techniques exploit the high temperature of fire compared to its surroundings. While this is a great use of IR video, people do not emit as much heat as fire and thus do not stand out as well. Also things that do not emit heat, such as backpacks or jackets which are clues in the search, are not very detectable in infrared.

Hajebi et al. [15] use stereo IR cameras and show that standard methods for doing image correlation do not work when using IR cameras. Lin [16] also talks about the many challenges of applying general computer vision techniques to infrared images. He notes that a few of the problems come from sensor types (as the IR sensors have slower response), resolution (IR is also lower res), and ranging problems as the spectra has a higher dynamic range. New methods need to be developed to successfully work with IR cameras.

1.1.3 Calibration of Two Cameras

Aligning multiple images or video streams is not a new concept. Calibrating two cameras for alignment of their videos has been done for a long time for use in stereo reconstruction [17, 18]. These methods are set up to use to Visible cameras whereas we want to work with one Visible and one IR camera.

Many researchers [19, 20] use manual methods for aligning multi-spectral images. They create a homography (mapping) using these manual methods, which allows them to combine the two images together. This method is dependent on the user's ability to manually detect similar points in each image to align them. It is not dependent on the types of cameras being used.

Mutual information [21] is another technique researchers have used to align images with different modalities. This process tries to find the best statistical explanation or mapping from one image to the other, minimizing the incorrect matches or joint entropy. This process has been deeply pursued in the area of medical imaging with images from different sensors such as MRI and CT [22]. Mutual information has also been looked at in aligning aerial video and reference imagery [23], since there can be great changes in an area that make standard alignment methods fail but where this type of statistical method can do well. The process however is computationally

intensive, so we are looking for something a little more lightweight that can work for our application.

In medical imaging, markers have long been used to align different image modalities. These markers can be physical material placed into or onto the patient or object being imaged to provide a reference in multiple modalities. With these markers on the patient, they can be transported between the various imaging devices to be used and the images can be later aligned. Maintz et al. [24] describe a number of ways these markers are employed. In our calibration methods we make use of this type of technique by using objects that are visible in our multi-modal images to calibrate our cameras, however we have a different alignment problem since our cameras are in a fixed configuration that can capture both images simultaneously allowing us to pre-calibrate the alignment.

Aligning Visible and IR imagery has also been done before. Irani and Anandan [25] use Laplacian energy images in aligning the two modalities. They compute a normalized cross correlation to determine the correctness of a current warp in a Lucas-Kanade [26] style iteration. Again the computational overhead for doing this with frames is prohibitive for live video display.

Researchers have also used silhouette extraction [27, 28] to align the frames based on the silhouette. This method does not provide a full perspective transform needed for our work.

1.1.4 Using IR and Visible video together

A number of researchers have worked with IR and video together to gain a better understanding of what they are looking at. Rudol [29] et al. use a combination of IR and Visible video to find human bodies in a search setting. They have to have correlated data to use these two images together. Rather than aligning their images, they use the pose information of the aircraft and the ground plane to map the location

from one camera to the real world and then back to the other camera, thus correlating a small area in the images. We want to be able to have a mapping without the need to have pose data for the UAV. We also want to display the information to a searcher as there are objects of interest that would not be able to be automatically detected using their methods.

1.1.5 Image Fusion for Display

Image fusion has been used for a number of purposes to assist users in understanding or simplifying information being presented. Researchers have used image fusion to provide better contextual views. Raskar et al. [30] use a stationary Visible camera to fuse daytime and nighttime imagery, producing a contextual nighttime view for observers. The task that they work with is much simpler than ours since anything that is visible in the nighttime imagery should be superimposed on the daytime imagery. In our research this would occlude information from one of the sensors and would negate the benefits of flying both at the same time.

Other researchers have combined different spectra to give the user a single view that combines the information from many sensors [31, 32]. This work has only used greyscale sensors, which simplifies the problem of which band or spectrum is picked to display. It allows the systems to merge similar features using averaging or selecting salient features where the different bands do not match. In Search and Rescue, the color information obtained from the Visible camera is very important and cannot be combined in the same way. A new method that can handle multi-channel color information needs to be developed.

1.2 Thesis Layout

The remainder of this thesis will be laid out as follows. We will first present a paper (Chapter 2) that we are submitting to The Twelfth IEEE International Conference on

Computer Vision (ICCV 2009). This paper will go over some of the same introductory material and related work we have already presented. It will go into detail on the methods we use to obtain our fused image, and show results produced by the methods. It also discusses the results found in a user study performed to validate the fusion methods. Chapter 3 goes into further detail on some of the methods that were not appropriate for the paper due to scope and space limitations. It also presents extended descriptions of a user study and the results obtained from this study. Chapter 4 will present our conclusions of this work and talk about future ways it could be enhanced and extended.

Chapter 2

Combined Visible and Infrared Video for use in Wilderness Search and Rescue

We now present a paper that has been submitted for pulication and is under review. The paper has had formatting changes to match the formatting of this thesis.

Abstract

Mini Unmanned Aerial Vehicles (mUAVs) have the potential to be a great asset to Wilderness Search and Rescue groups by providing a bird's eye view of the search area. These vehicles can carry a variety of sensors to better understand the world below. This paper proposes using both Infrared (IR) and Visible Spectrum cameras on a mUAV for Wilderness Search and Rescue. It details a method for combining the color and heat information from these two cameras into a single fused display to reduce needed screen space for remote field use. To align the video frames for fusion, a method for simultaneously pre-calibrating the intrinsic and extrinsic parameters of the cameras and their mount using a single multi-spectral calibration rig is also presented. A user study conducted to validate the proposed image fusion methods showed no reduction in performance when detecting objects of interest in the single-screen fused display compared to a side-by-side display. Furthermore, the users' increased performance on a simultaneous auditory task showed that their cognitive load was reduced when using the fused display.

2.1 Introduction

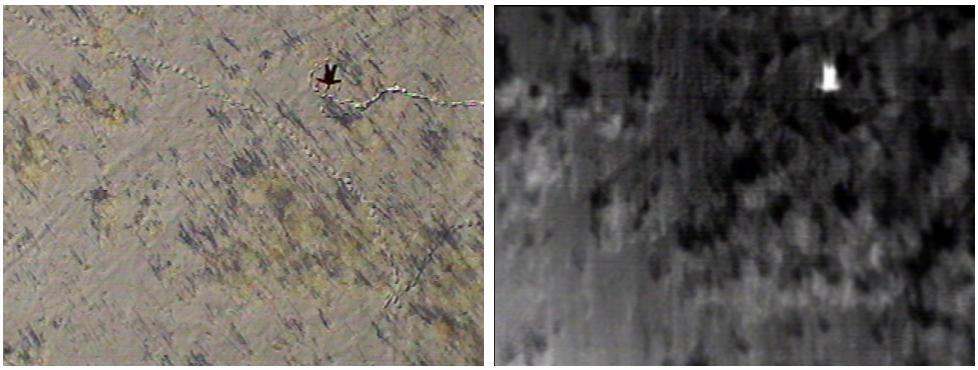
Search and Rescue teams throughout the Western United States are often called out to assist individuals who find themselves lost or in peril. These teams often utilize pilots to assist them from airplanes or helicopters, giving them a bird's eye view of the area. These aerial searchers speed up the search by covering large areas in a short period of time.

Aerial searching, however, comes with a number of disadvantages. Conventional aircraft are very costly to purchase and operate and pose potential danger to the pilot and crew members. In 2006 there was an incident in Utah where a sheriff's deputy was killed during a search when the helicopter he was flying in hit power lines and crashed [1].

In recent years, there have been great advances in the use of Unmanned Aerial Vehicles (UAVs) to obtain video from the air. This aerial video can replace aerial searchers and reduce the costs and dangers of searching from the air. Mini Unmanned Aerial Vehicles (mUAVs) have also been used because they bring additional advantages. mUAVs can be launched on location in many terrains and cost far less to purchase and operate, making them more attainable for Search and Rescue groups that have limited funding. Throughout the remainder of the paper, when we use the term UAV we are specifically referring to mini UAVs.

UAVs are capable of carrying a variety of sensors. The most common sensor carried by these is a visible-spectrum color camera, which we will refer to as a "Visible camera". These cameras provide views similar to those that pilots can obtain in manned aircraft.

Another sensor that has potential to be very helpful in Wilderness Search and Rescue is an Infrared (IR) camera. This type of camera increases the information that an aerial searcher can obtain by being able to see the heat our bodies produce. IR cameras are most useful during times or in areas when emitted body heat is



(a) Visible (b) Infrared

Figure 2.1: Aligned and synchronized frames from Visible and IR cameras showing a person lying on the ground.

greater than heat emitted from the surroundings, such as during the nighttime, in the early morning hours, over snow covered ground, or over water. In many of these conditions, IR cameras could be used in conjunction with Visible cameras to maximize the information that the searchers receive.

Searching with both IR and Visible cameras presents a great deal of information to users all at once. Figure 2.1 shows an example of the information from aligned Visible and IR frames. Each frame, from both cameras, has the potential of containing new information to be analyzed, yet the viewer only has a fraction of a second to do so. Significant screen real estate is needed to display both videos simultaneously, making a small portable system for use in Wilderness Search and Rescue difficult to build and use.

A UAV equipped with IR and Visible cameras has the potential to be an excellent asset to a Search and Rescue team. This paper presents a novel method for overcoming the increased information and screen space needed to use these two cameras. This is done by combining the information from the two cameras into a single view, retaining enough information from both image modalities so that searchers are able to successfully find people and objects of interest. To be capable of combining the information we have also developed a new method for calibrating the intrinsic

and extrinsic parameters of our multi-modal cameras and their shared mount using a multi-spectral calibration rig.

2.2 Related Work

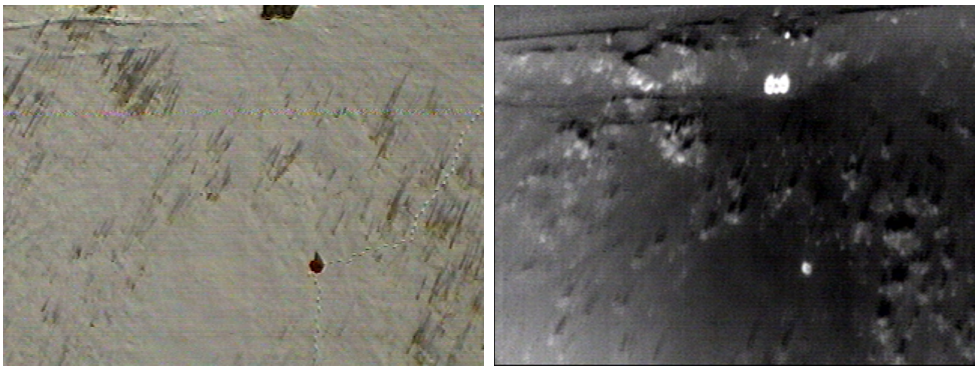
The area of aerial search and surveillance using UAVs has seen an expansion in recent years due to their increasing availability (for example [2, 4, 33]). This expansion has opened the way for different areas of research using aerial platforms for obtaining video.

Aligning multiple images or video streams is not a new concept. Some have used manually-selected points to align frames from different modalities [19, 20]. Automatic calibration of two cameras for alignment of their videos has been done in stereo reconstruction [17, 18]; however, these methods don't often work well when working with different image modalities.

Mutual information [21, 22, 23] has been shown to be an effective method for working with different modalities when images have enough statistical correlation, but it requires significant computation. Silhouette extraction [28, 27] has been used for IR and Visible alignment in systems where a person is walking across the scene, but this requires a specific object that can be singled out in both modalities and only aligns the extracted object rather than the entire frame.

Image markers have been employed in medical imaging to align different modalities [24, 34] but can only be used when the markers can be set up in advance. While we cannot set up markers on the ground in all areas where searches will be performed, we can use external markers to pre-calibrate the camera mount and align the images.

Once the images have been aligned they can then be interpreted. Many researchers work with IR and Visible imagery to make decisions on whether there is fire [13], or whether a person is in view [28, 29] These methods look at each sensor independently and use the correlation of objects in both frames to make decisions.



(a) Visible (b) Infrared

Figure 2.2: An example of the raw video from the cameras. The frames are synchronized, but not aligned.

In Wilderness Search and Rescue there are many objects we look for that may not show up simultaneously in both modalities, so a method to present both forms of this information to the user is needed.

Many have worked in the area of image fusion [31, 32], visually fusing information from two or more modalities into a single view. These researchers have worked with greyscale sensors, which allow them to find features from the different images to include in their output image. In Wilderness Search and Rescue, the color information obtained from a Visible camera is important and cannot be combined in the same way that greyscale images can. Rather than a single channel of information from each sensor, we have multiple channels from the Visible camera that needs to be looked at jointly to retain the color information.

2.3 Methods

To combine video from IR and Visible cameras, the individual cameras as well as their shared mount need to be calibrated. Once the calibration has been performed, the frames can be warped into alignment and combined into a single image. Figure 2.2 shows sample frames before any alignment has been performed.

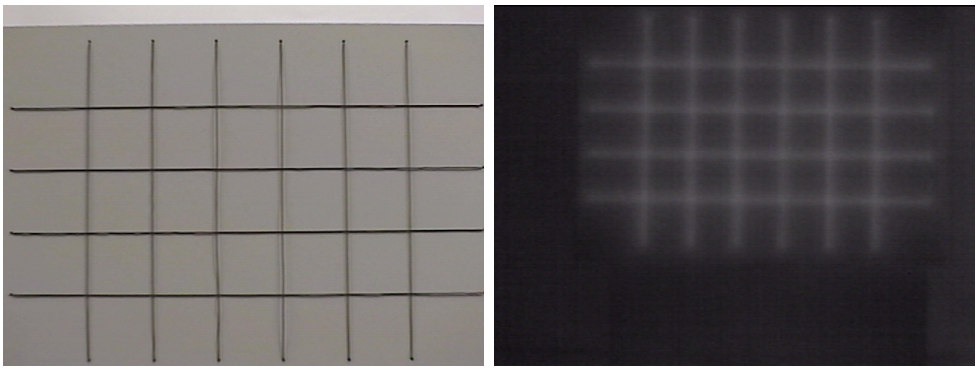
2.3.1 Image Alignment

To align the IR and Visible imagery, we need to first calibrate each of the cameras, followed by calibrating the mount that holds the cameras together. These could be done separately; however, we have developed a method that can do both of these in a sequential fashion using the same data. To do this, a set of objects need to be chosen that can provide both internal camera calibration as well as external mount calibration. To be useful, these objects need to be detectable in both the Visible and Infrared spectra. The camera calibration requires a pattern with known distances, and the rig calibration requires some way of pairing corresponding points from one image to the other. To satisfy all of these constraints simultaneously we use a grid pattern of wires (Figure 2.3) that have a small electrical current running through them so that they warm up slightly and emit heat. From this grid, corners are extracted where the wires meet, and then these points are used to calibrate the intrinsic and extrinsic parameters of the cameras and their mount.

Point Extraction

Extracting the points from the grid of lines (Figure 2.3) is done by first using the Hough transform [35] to identify possible lines in the image. Intersecting lines are checked to verify that the intersection has at least a 45° angle between them (to remove near parallel intersections), and these lines are intersected to produce possible points. These points are then fit to the grid pattern we are trying to recover, and checks are performed to verify that the recovered pattern is a grid. The user can easily reject any bad set of points that may make it through all of the checks.

Once the grid is located, the points are refined using the lines that define them. A small window is extracted around each point to perform the refinement. We cannot use the entire line, as this would remove any distortions due to the lens and drastically



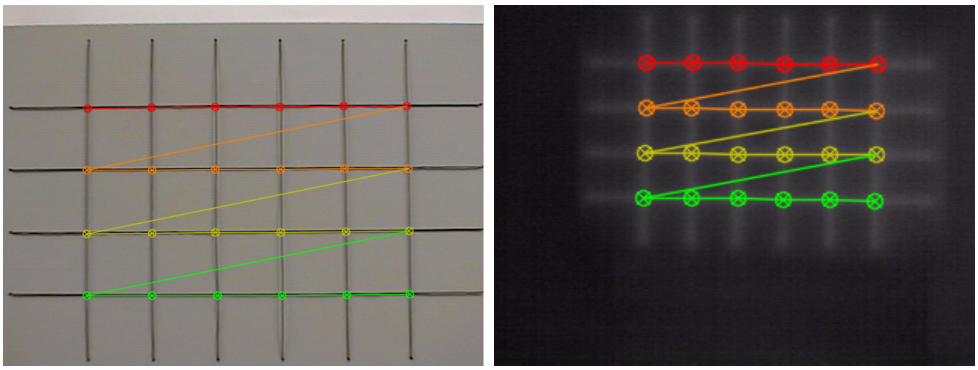
(a) Visible (b) Infrared

Figure 2.3: The multi-spectral calibration rig.

change or skew our calibration results. By using a local area the distortions remain intact and can be correctly discovered and removed in the camera calibration.

To refine the points, we first estimate the equation of each line, then determine the intersection of these lines. Weighted linear regression is used to refine the line parameters for each line to reduce inaccuracies in the corner locations introduced by the Hough transform as well as from the limited resolution and sensitivity of the cameras (as shown in the fuzziness of the lines in Figure 2.3).

Since two separate lines are present in the small window around the intersection, not all of the points can be used in the refinement. A small area around each line can be used, but to do so requires an approximation for the line. Due to the grid nature of the points, an approximation of the line can be found by using the neighboring intersections. An E-M [36] style iteration is then used in which the line approximation is updated and reapplied to the data to remove the biasing due to the original line approximation. The refinement provides much more accurate point locations and produces good results in the calibration. Figure 2.4 shows the results of finding and refining the point locations.



(a) Visible (b) Infrared

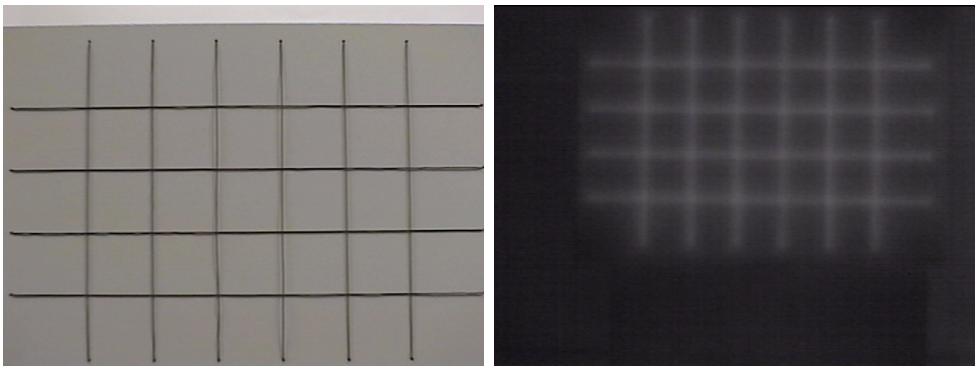
Figure 2.4: Located grid points after applying refinement.

Intrinsic Camera Calibration

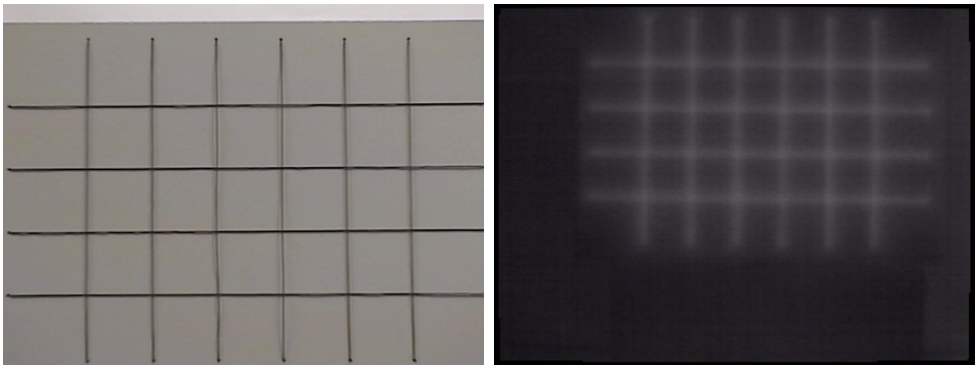
To calibrate each camera, the grid points from each acquired image are used. The Bouguet method for camera calibration is then applied to the image points [37, 38]. This method gives both the intrinsic camera parameters as well as the distortion coefficients that correct for lens distortions. Figure 2.5 shows the lines after camera calibration has been performed and the distortions removed.

Extrinsic Camera Calibration

To calibrate the extrinsic parameters of the camera mount, the translation and rotation between the cameras must be recovered. Rather than recover these individually we recover a homography between the frames that incorporates both of these values into one mapping from the IR to Visible frame. A homography is sufficient for our setup since we need to fly between 60 and 100 meters above the ground, and we can treat each image as if it is of a planar scene. This is done with the same set of points used to calibrate the intrinsic parameters of the cameras. Due to properties of a homography, the translation between the cameras affects the homography at the distance used to capture the points for calibration, but when the cameras are on a UAV high above the ground this translation is negligible and the homography



(a) Original Visible (b) Original Infrared



(c) Undistorted Visible (d) Undistorted Infrared

Figure 2.5: The same lines as shown in Figure 2.3 before and after intrinsic calibration and application of the distortion parameters. It is easiest to notice in the Visible images that the lines are straighter after performing the calibration and removing the lens distortions.

reduces to a rotation. The rotation between the cameras is obtained by decomposing the homographies found during calibration and removing the translation component.

To find the homography, multiple image pairs are used and the results are averaged. The homography from each image pair is taken separately due to changes in the orientation of the grid plane that modify the plane normal and resulting translational component of the homography. The set of point pairs p_i and q_i from the Visible and IR images respectively are brought into a similar world reference frame by applying their respective calibrations matrices K_1 and K_2:

$$p_i' = K_1^{-1} p_i \tag{2.1}$$

$$q'_p = K_2^{-1} q_i \tag{2.2}$$

The points p'_i and q'_i are then used to calculate the homography H such that $p'_i = Hq'_i$ using the 8 point algorithm [39]. H is then decomposed into rotation R, translation $\frac{1}{d}T$, and plane normal N as in [40]:

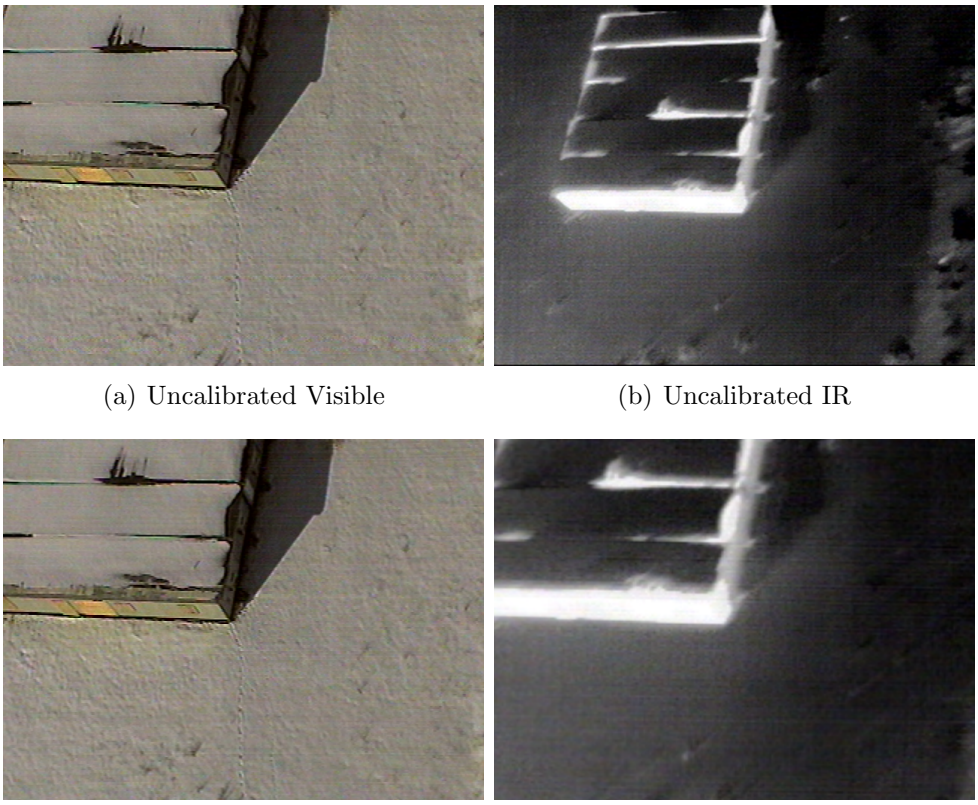
$$H = R + \frac{1}{d} T^T N \tag{2.3}$$

This decomposition produces four possible solutions that correctly recreate the homography, however only one is physically correct. Two of the solutions can be immediately removed since the z component of their plane normals is negative, orienting the plane facing away from the cameras. We remove the final ambiguity by again looking at the z component of the plane normal N and choosing the solution with the largest z component, since the plane on which the points lie is always close to perpendicular to our cameras.

Once each rotation has been isolated, the set of rotations are averaged together to get the final rotation estimate. This rotation then needs to be brought back into the image space, which will add in any scale and possible translational differences due to the cameras. This is done by applying the calibration matrices (K_1 and K_2) from the cameras to get the warp W that aligns the IR frame to the Visible frame:

$$W = K_1 R K_2^{-1} \tag{2.4}$$

The results of the extrinsic camera calibration are shown in Figure 2.6 and 2.7. The misalignments are easy to see in the uncalibrated images of both figures but have been removed through the calibration methods as shown in the calibrated images. In Figure 2.7, the Visible and Infrared images are displayed in a checkerboard fashion, showing how well the edges of the building are aligned in the two images.



(a) Uncalibrated Visible (b) Uncalibrated IR

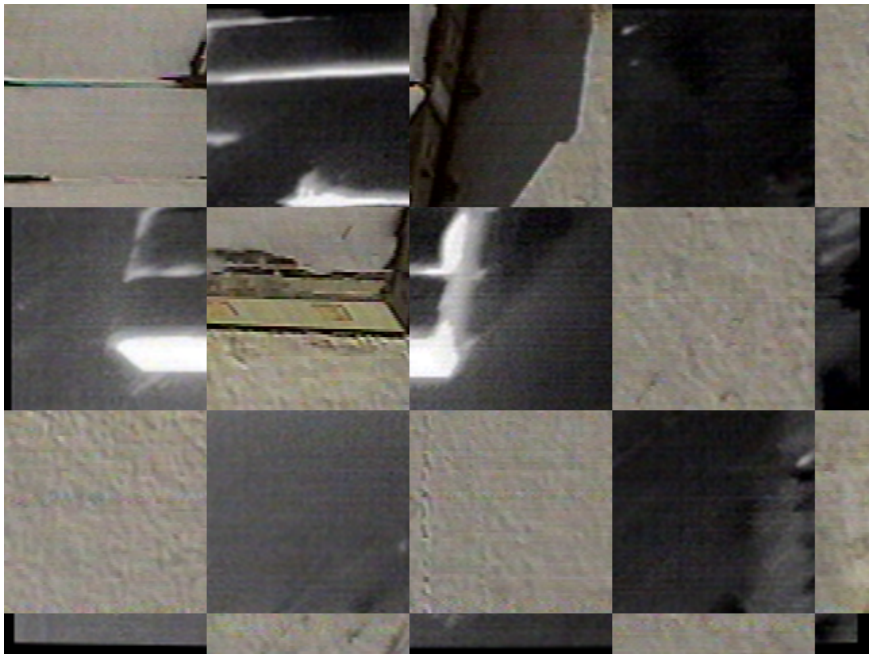
(c) Calibrated Visible (d) Calibrated IR

Figure 2.6: Sample image pair before and after calibration.

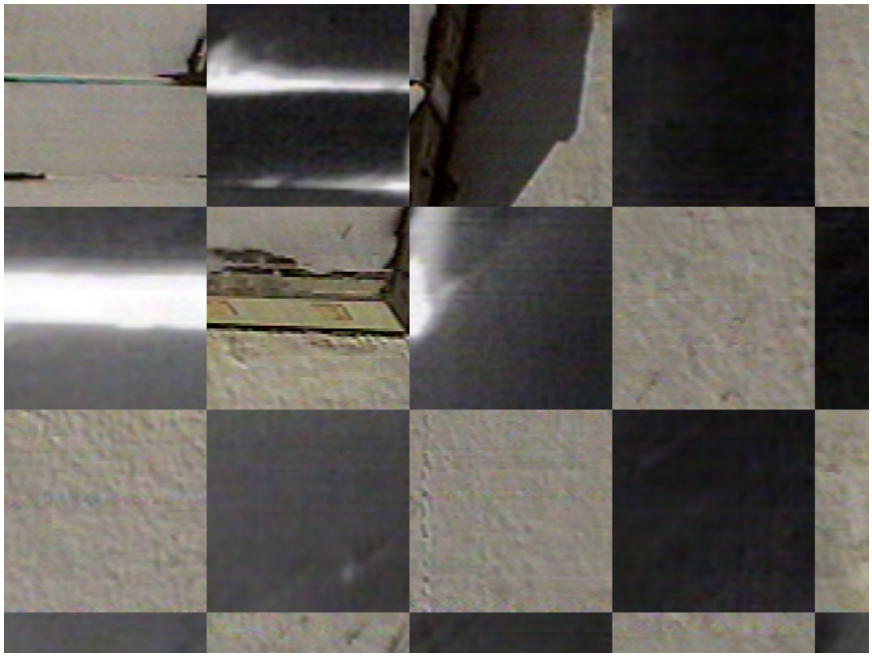
2.3.2 Image Fusion

Combining the IR and Visible imagery into a single image has the potential for greatly reducing the load on the user watching the video, as well as saving available screen space for field use. Our method involves highlighting objects (or areas) in the Visible frame where hot objects are found in the Infrared frame by creating a heat overlay to place over the Visible image. This is done by taking advantage of colors that are not common in the wilderness areas we image and also of the temporal dimension of the video.

To perform this highlighting we start with images that are aligned using the calibration methods described in Section 2.3.1. We create an overlay image $O(x, y)$ containing the heat information to be shown to the user. An HSV value is assigned



(a) Uncalibrated



(b) Calibrated

Figure 2.7: Checkerboard composite examples of the calibration using the same image frames from Figure 2.6. These images show both the IR and Visible images displayed in a checkerboard fashion on top of each other to show the alignment. (a) shows the alignment before calibration and (b) shows the alignment after calibration.

to our overlay depending on whether the value of the IR image $I(x, y)$ is above a specified threshold t:

$$O(x,y) = \begin{cases} HSV(H, I(x,y), I(x,y)) & \text{if } I(x,y) > t \\ HSV(0,0,0) & \text{otherwise} \end{cases} \tag{2.5}$$

When the IR value is above the threshold we give the overlay image a pre-selected hue H and the IR value $I(x, y)$ for both its saturation and value. This allows the color to be brighter or darker when the object is hotter or colder. For all of our video we picked a hue of 300° on a 360° color wheel, giving us a magenta color. This is a rare color in wilderness areas and it helps give the user a sense of heat at the same time. Any hue can be used that gives the searcher an understanding of what is hot and is atypical in the target search area. We use a threshold of 150 for segmenting the heat information in the infrared video, though this can be adjusted by the user. The effect of the threshold on detection rates is left as an area of future work.

After the overlay is created we combine it with the Visible image $V(x, y)$ to create our fused image $F(x, y)$. The threshold t from Equation 2.5 is used again. A user-controllable opacity α is used to overlay $O(x, y)$ on $V(x, y)$ to produce the fused image $F(x, y)$:

$$F(x,y) = \begin{cases} V(x,y)(1-\alpha) + O(x,y)\alpha & \text{if } I(x,y) > t \\ V(x,y) & \text{otherwise} \end{cases} \tag{2.6}$$

The opacity allows the user to control to what extent they see the original Visible image vs. the thermal overlay, tailoring the output image to the user's preferences.

The colored overlay has the potential of either reducing the understanding of the original Visible image due to the added color or to be barely noticeable. To compensate for this, the temporal nature of video is leveraged by turning the overlay on and off at a user-specified rate (since anecdotal evidence suggests that users find

(a) Visible (b) IR (c) Fused

Figure 2.8: Example of IR and Visible image fusion. This result uses 50% transparency.

different rates to be effective), thus interleaving a number of original Visible frames with the new fused frames. This allows the original information to be visible for a few frames and then the heat information to also be available (overlaid) for a few frames. This also attracts the visual attention [41] of the user.

Figure 2.8 shows an example of this method. We can see the original Visible and Infrared frames that were used as well as the resulting fused frame. This image was created using 50% opacity, allowing the original Visible frame to still be seen through the overlay.

2.4 Results

All of the video and images in this paper were obtained using a KX141 color camera from Black Widow AV [42] and FLIR's Pathfind IR infrared camera [43]. The Visible video was captured at 640×480, and the Infrared video was captured at 320×240. The aerial video was obtained by flying the cameras on a five-foot flying-wing remote control plane (Figure 2.9a) outfitted with the Procerus Technologies Kestrel Autopilot system [44].

A simple camera mount (Figure 2.9b) was built to retain the camera configuration, and the cameras and mount were calibrated using the methods from Section 2.3.1.

(a) Mini UAV (b) Cameras and Mount

Figure 2.9: The hardware used to obtain the aerial imagery.

Figure 2.10 shows a sequence of frames using our fusion methods including the Visible frame, the aligned Infrared frame, and our fused frame. The frames show the image of a person lying on the ground passing though the frames.

To validate our image fusion method, we conducted a user study to compare performance on a detection task given Side-By-Side Visible and Infrared videos compared with the fused video (which was labeled "Combined" in the study). To test the subjects' cognitive load while performing this task, a secondary task was added, in which they were asked to count the number of tones that were played through headphones. This secondary task had a low- and a high-difficulty setting to adjust the cognitive load placed on the user. In the low-difficulty setting, subjects were asked to count how many times a single tone was played. In the high-difficulty setting, they were asked to count two different tones and keep track of how many times each was played. This secondary audio task loosely simulates what a searcher may need to do as they interact with others who control the plane or need information about objects found while still searching through new video. Performance on this secondary task reflects the amount of mental workload required in the primary task; high performance means low workload and vice versa.

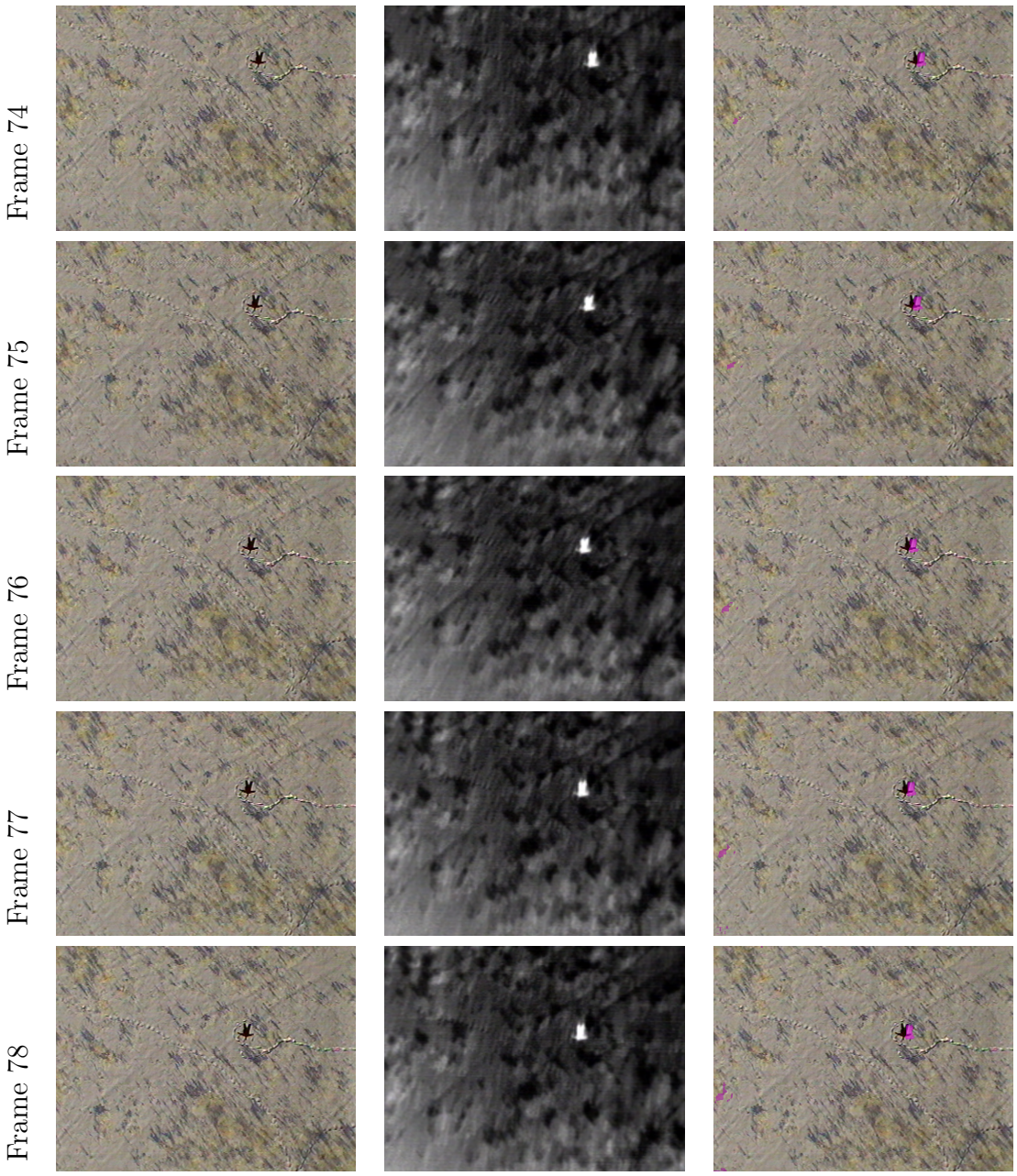


Figure 2.10: An image sequence showing a person lying on the ground.

The user study involved 32 volunteer subjects. Eight videos were watched by each subject, which collectively contained 15 objects: eight people lying on the ground and seven red circles. The videos were taken during the winter, and the objects were placed to try to require using information from both the IR and Visible videos for detection. The video order, display methods, and secondary tasks were randomized

Display Method	Least Squares Means (Std. Error) of False Positives
Combined	1.8454(0.3033)
Side-By-Side	2.4436(0.3033)

Table 2.1: False positive results from the user study. The difference has a p value of 0.1516. While this is not extremely significant it was the next closest significant item in all of the statistics run against the primary task.

to minimize ordering effects, and subjects saw each combination of display method and secondary task twice. We analyzed the data using mixed models ANOVA with subjects as a random blocking factor. All other variables were fixed effects and their differences were evaluated. The results are presented as Least Squared Means.

On the primary task, subjects were able to detect 90% of the objects without any significant differences due to the display method or the secondary task being performed. The relative difficulty of the individual videos was the only statistically significant ($p < 0.0001$) factor that affected the performance on this task. There was a slight statistical trend ($p = 0.1516$) shown in the number of false positives subjects detected depending on the display method, where there were approximately 25% fewer false positives using the Combined display (Table 2.1).

On the secondary task, a clear improvement came when using the Combined display. Subjects reported the number of tones played with approximately 33% better accuracy ($p = 0.0417$) while viewing the Combined display than while viewing the Side-By-Side display (Table 2.2). This strongly suggests a reduced cognitive load when working with the Combined display.

The subjective feedback confirmed the findings of the primary and secondary task analysis. On the preference questions asking which display was preferred for the detection task and which display would be preferred for use in a real search and rescue task, the responses were mixed, confirming the primary analysis that showed similar detection performance with either the Combined or Side-By-Side display. On the

Display Method	Least Squares Means (Std. Error) of Miscounted Tones
Combined	0.9067(0.1808)
Side-By-Side	1.3511(0.1808)

Table 2.2: Miscounted tones from the user study. The users more accurately reported the number of tones played when watching the Combined video display compared with the Side-By-Side video display. The difference has a p value of 0.0417.

question asking which display was easier to watch, more subjects felt that the Combined display was easier, confirming our secondary task findings that the cognitive load was less with this display.

2.5 Conclusions

Using our calibration and fusion methods we are able to create a fused display that allows users to do just as well as they can with a simple side-by-side display, while requiring less cognitive effort on the part of the users and less screen space. This has great potential for assisting searchers when using both of these imaging modalities simultaneously.

The multi-modal video alignment developed here to align the frames for image fusion opens up a number of different areas that can be pursued with both IR and Visible cameras. Infrared video mosaics can easily be created using frame alignment information from the Visible video. With this same video alignment, filtering and super-resolution could be performed on the Infrared imagery to provide better information to the user.

The fusion methods we developed could be enhanced. Due to the thresholding of the infrared video, information is lost that could be detectable when looking at the separate videos. This might be mitigated by using adaptive thresholding or a segmentation method suitable for Infrared imagery. Our fusion method has not been

tested in all settings. In our user study and in all of the aerial images presented in this paper we used imagery of winter scenes with snow on the ground. The fusion methods need to be tested during the summer, and some adaptation of the thresholding may need to take place to correctly segment a person's heat vs. the heat of other objects. Fatigue levels would also be interesting to test using our fusion methods compared with using side-by-side videos, though with the decreased cognitive load we could expect that the fatigue levels would be significantly reduced when using the fused display.

The calibration and fusion methods developed in this paper show great potential for assisting users in Wilderness Search and Rescue. They allow heat information to be added to the color imagery obtained from UAVs. The fusion methods decrease the cognitive load on the searcher while maintaining the ability to correctly detect objects of interest.

Chapter 3

Additional Details

This chapter goes into further detail on some of the methods that were not appropriate for the paper due to scope and space limitations. It also presents extended descriptions of the user study performed to validate the fusion methods and the results obtained from this study.

3.1 Weighted Linear Regression

To perform accurate calibration requires that point estimates be as accurate as possible, which can be done by applying weighted linear regression. Weighted linear regression is used to refine the line parameters for each line to get the best estimate possible. These lines are then intersected producing accurate point estimates to be used in the calibration methods.

To perform weighted linear regression we parameterize the line into β and ϵ as follows:

$$y = \beta_y x + \epsilon_y, \qquad x = \beta_x y + \epsilon_x \tag{3.1}$$

These two different formulations are used instead of a single one due to problems when working with horizontal and vertical lines. The problem can be removed by picking the formulation that has the largest β as calculated in Equations 3.2 and 3.3.

To calculate the values for β and ϵ the following equations are used:

$$\beta_y = \frac{\sum_{i=1}^{n} w(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} w(x_i - \bar{x})^2}, \qquad \epsilon_y = \bar{y} - \beta_y \bar{x} \tag{3.2}$$

$$\beta_x = \frac{\sum_{i=1}^{n} w(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} w(y_i - \bar{y})^2}, \qquad \epsilon_x = \bar{x} - \beta_x \bar{y} \tag{3.3}$$

where $\bar{x}$ and $\bar{y}$ are the average x and y values from the observed points. The weight term w is taken from the intensity of the greyscale image.

3.2 Intrinsic Camera Calibration

The method used for determining the internal camera calibration can be found in both a Matlab Toolkit [45] as well as in OpenCV [46]. These also provide methods for removing the distortions caused by the lens that are used in this work.

3.3 Homography Decomposition

A homography is a 3×3 matrix that warps each point in an image plane to a different plane. It is often used in image registration to define the warp from one image to another:

$$x_1 = H x_2 \tag{3.4}$$

Homography matrices are made up of 4 components: rotation R, translation T, plane normal N and distance $\frac{1}{d}$ as shown in Equation 2.3

Michaelson [40] describes the method to take a given homography and transform it back into its components. There are a few ambiguities that cannot be completely corrected for without further information. The first of these is that the distance and the translation cannot be separated. Michaelson's derivation therefore incorporates this into a single value t removing the distance portion. The second ambiguity

results in four possible solutions that require external data to resolve. We will first show the steps to perform the derivation as described in [40] and then show how to reduce the possible solutions to the correct solution given our setup.

To perform the decomposition, we first apply singular value decomposition to get $H = UH'V$. We then transform Equation 2.3 and get $H' = R' + t'n'^T$ where $R = UR'V^T$, $t = Ut'$ and $N = Vn'$. The singular values are h_1, h_2 and h_3. We scale h_2 to equal 1 to simplify:

$$\begin{pmatrix} h_1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & h_3 \end{pmatrix} = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix} - \begin{pmatrix} t'_x n'_x & 0 & t'_x n'_z \\ 0 & 1 & 0 \\ t'_z n'_x & 0 & t'_z n'_z \end{pmatrix} \tag{3.5}$$

This can be taken a step further to get

$$n' = \begin{pmatrix} s_1\sqrt{\frac{h_1^2-1}{h_1^2-h_3^2}} \\ 0 \\ s_2\sqrt{\frac{1-h_3^2}{h_1^2-h_3^2}} \end{pmatrix}, \quad t' = (h_1 - h_3)\begin{pmatrix} n'_x \\ 0 \\ n'_z \end{pmatrix}, \quad \sin\beta = (h_1 - h_3)n'_x n'_z \tag{3.6}$$

where $s_1 = \pm 1$ and $s_2 = \pm 1$, yielding four possible solutions.

To reduce the four possible solutions to the single correct solution for our setup, we first remove two of the solutions that are not possible since the z component of the plane normal N is negative. This would put the imaged objects behind the camera, which is not possible. From the remaining two solutions, the correct solution is found by checking the plane normal N and finding the solution with the largest z component. This works in our setup since the plane on which the points lie is close to perpendicular to the camera's optical axis.

3.4 User Study Details

To perform the user study we obtained 8 videos to display to the users. These videos were taken during the winter. Each of these videos was taken using our camera mount that had been previously calibrated. The videos collectively have a spread of 15 different targets: eight people lying on the ground and seven red circles.

The objects in the video were placed to try to make the user utilize information from both the Visible and Infrared video to locate them. The largest variations we were able to obtain were seen in how the people appeared. The person targets were either a person lying on the ground under a white blanket (to mask their detectability in the Visible video), a dark shirt and pants lying on the ground that had heated up due to the sun's infrared radiation, or a dark shirt and pants that was cooled off so as to not emit heat. With the red circles, they were always detectable in the Visible video because their color was what set them apart. They were not however always detectable in the infrared as some of them were hot due to the sun's infrared radiation while others were not.

The study consisted of two different tasks that were performed simultaneously. The primary task was to watch the video and indicate when a person or red circle was seen. The secondary task was to count how many tones were played through headphones and indicate the count after the video was finished.

For the primary task, the user would see each of the 8 different videos. Each video was shown in either a separated side-by-side view (Figure 3.1) with the Visible video on the left and the Infrared on the right, or a combined view (Figure 3.2) using the fusion methods described in Section 2.3.2. For the fused presentation we used a threshold of 150 and a transparency value of 50%. The user was asked to hit the '**z**' key when they saw a person and either the '**x**' or the '/' key when they saw a red circle (both keys were given as an option to allow for the user to decide if using two fingers or two hands was easiest). Small stickers were placed on the keys to help the user

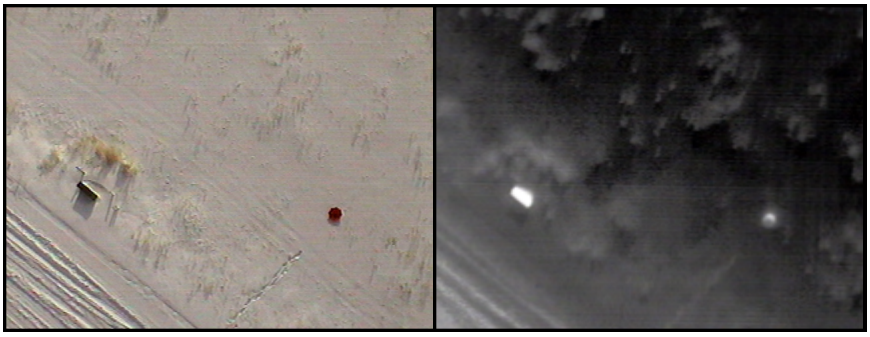


Figure 3.1: A still frame of the side-by-side view. This frame has a red circle in it.

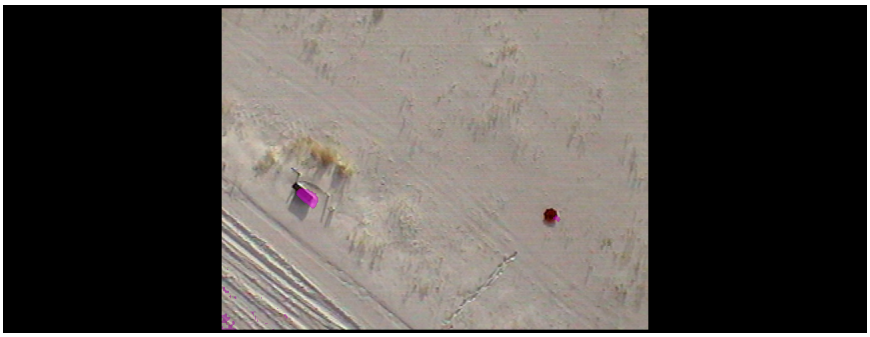


Figure 3.2: A still frame of the combined view. This frame is the same as the one shown in Figure 3.1.

easily identify which keys were to be used and to reduce the problem of remembering which key to use.

For the secondary task, the user was asked to count a number of tones. This task was used to add a cognitive load to the users. There were two different difficulty levels in this task: the first required the user to count how many times they heard a high pitch tone, the second required the user to count how many times they heard a high pitch tone and how many times they heard a low pitch tone. The number of tones, frequency of occurrence, and order of the tones were all random.

For the primary task, each key press of the user was recorded and later analyzed to determine the number of people or red circles they correctly marked, the number of people or red circles they missed, and the number of false positives on either people or red circles. The user was also asked how many objects they felt they saw but did not mark and to indicate how confident they were that they marked all of the

	1	2	3	4	5	6	7	8
0	SH	SL	CH	CL	SH	SL	CH	CL
1	SL	CH	CL	SH	SL	CH	CL	SH
2	CH	CL	SH	SL	CH	CL	SH	SL
3	CL	SH	SL	CH	CL	SH	SL	CH

Table 3.1: Lookup table used to assign display method and task for a given video.

objects in the video. For the secondary task, the user was asked to report how many tones they counted (for the two tone task they were asked how many of each of the tones separately). They were also asked to report how confident they were that they counted all of the tones.

To ensure that problems were not caused by the ordering, the videos, display methods, and tone counting tasks were randomized and stored in a file for each run. This was done by randomizing the order of the videos using a random number generator and making sure that each video came up once per person. The video display methods tasks were then assigned to each video. This was done by using a lookup table (Table 3.1) that ensured that each video display method and task was seen twice by each user. This table also made sure that each video, method, and task combination was seen once for every four people who performed the study.

The lookup table was used by taking the n^{th} run of the randomization and taking n modulus 4 to find the row of the lookup table. Then the video number was used to index to the column in the lookup table. The codes in the lookup table are 'S' for Side-By-Side video and 'C' for Combined or Fused display. The 'H' was for the two-tone task and the 'L' was for the single-tone task. The randomization of the videos took care of making the tasks and display methods occur in random order as well.

The user study included a training section that allowed the users to perform the same tasks on sample videos. For this section two more videos were obtained that were different from the eight used for the study but with the same types of targets

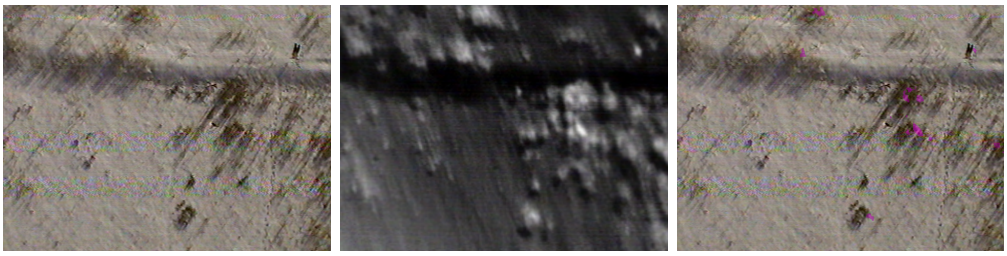
and during the same time of year. The user was allowed to repeat the training section as many times as they wished. The ordering and presentation of the training section was all randomized, but the user did see each video, display method, and secondary task once each time through the training.

When each user came to perform the study, they were given an instruction sheet explaining the tasks of the user study (Appendix A). The user was also given a Pre-Training Questionnaire which asked demographic as well as prior knowledge questions (Appendix B). Once the user was done with the video tasks, they were given the Follow-up Questions (Appendix C), where we asked them which of the two display methods they preferred and for their comments.

3.5 User Study Results

Our user study gave us a number of different metrics that allowed us to analyze our fusion methods. The first analysis was done directly on the primary task. We then analyzed the secondary task as well as the subjective responses. We analyzed the data using mixed models ANOVA with subjects as a random blocking factor. All other variables were fixed effects and their differences were evaluated. The results are presented as Least Squared Means.

In the primary task we did not see significant differences between the two methods. There was great statistical significance depending on which video the subject was watching ($p =< 0.0001$). Each video had a different area it searched through, which caused different difficulty levels depending on the video. We also had two objects that were regularly missed when viewed with both methods. These objects are shown in Figures 3.3 and 3.4. The person in Figure 3.3 was difficult to see because they were visible for only two-thirds of a second. The person in Figure 3.4 was visible for a much longer time but was cold, and there was a red circle that quickly appeared in the same frames, distracting many subjects. When the data from these two ob-



(a) Visible (b) IR (c) Combined (Fused)

Figure 3.3: Frames showing the person in Video 4 that was regularly missed.

(a) Visible (b) IR (c) Combined (Fused)

Figure 3.4: Frames showing the person in Video 5 that was regularly missed.

jects is ignored, only 2% of the rest of the possible objects were missed. These results show us that there is no significant difference in the subjects' performance on the two displays.

When analyzing the primary task, the next most statistically significant ($p = 0.1516$) difference was the display method which showed up when looking at the number of false positives the subjects marked. Table 2.1 shows the Least Squares Means analysis of the subjects' number of false positives indicated. For the Combined display the subjects had 25% fewer false positives than they did when using the Side-By-Side display. The statistical significance is not high enough however to draw any conclusions from this result.

On the secondary task we were able to draw some more significant conclusions. We analyzed the number of miscounted tones and found there there was a 33% decrease in the number of tones missed when using the Combined display compared to the Side-By-Side display ($p = 0.0417$). Table 2.2 shows the Least Squares Means

analysis for the number of miscounted tones. The increase in performance with the Combined display on this task shows that the subjects had a lower cognitive load when using that display method. This would indicate that this display method was easier to watch.

Finally we focused our analysis on the subjective questions that the subjects were asked throughout the user study. On the primary task, the subjects' responses showed that their confidence level for clicking on all of the targets was affected by which video they had just watched. The responses also showed a slight training effect in their confidence as well as in their indication of how many objects they felt that they missed. Our training section was set up to minimize this, but some effects still appeared in the confidence levels the subjects reported. Interestingly though, there was no significant effect that the ordering showed up in the actual performance, indicating the training effect was not apparent in the subjects' performance. Their subjective responses do support the results from the analysis on their performance on this primary task.

The confidence levels of the subjects on the secondary task was most significantly affected ($p = 0.0513$) by the interaction of the display method and the level of difficulty on the secondary task. Table 3.2 shows the results of a Least Squares Means analysis. When looking at the confidence values, it is important to note that zero indicated the highest confidence and four was the lowest confidence. It appears that subjects did the best when working with the Combined display on the single-tone secondary task. The interesting interaction is that the subjects confidence increased when going from the more difficult task to the easier task when using the Combined display, but it increased when looking at the Side-By-Side display. We are not sure what would have caused this and feel that it is probably an anomaly in the data. On this same data the next most statistically significant differences ($p = 0.0938$) appeared when looking specifically at the display methods (Table 3.3), with higher

Display Method	Secondary Task	Least Squares Means (Std. Error) of Tone Counting Confidence
Combined	**Two Tone**	1.7022(0.1921)
Combined	**Single Tone**	1.5729(0.1925)
Side-By-Side	**Two Tone**	1.6733(0.1926)
Side-By-Side	**Single Tone**	1.9109(0.1927)

Table 3.2: Confidence levels reported by subjects in the User Study when comparing the interaction of the display method as well as the secondary task difficulty. Zero was the highest confidence level while four was the lowest. The differences shown in this table had a P value of 0.0513.

Display Method	Least Squares Means (Std. Error) of Tone Counting Confidence
Combined	1.6376(0.1359)
Side-By-Side	1.7921(0.1359)

Table 3.3: Confidence levels reported by subjects in the User Study when only the display method is compared. Zero was the highest confidence level while four was the lowest. The differences shown in this table had a P value of 0.0938.

confidence reported using the Combined display. This was not significant enough to draw conclusions from but does support the results seen when analyzing the number of miscounted tones.

The final subjective questions analyzed are those found on the Follow-Up Questionnaire (Appendix C). In the first question regarding the effect of the misalignments, only 25% of the subjects felt that the infrequent misalignments caused difficulty when watching the fused video. This is important to note that for some people their performance on the fused display may have been better if no misalignments were present.

Response	Percent
Side-By-Side was easiest	18.75%
Side-By-Side was slightly easier	31.25%
There was no difference	3.125%
Combined was slightly easier	28.125%
Combined was easiest	18.75%

Table 3.4: Subject responses to which display method was easier to find objects in.

Response	Percent
Side-By-Side was easiest	12.5%
Side-By-Side was slightly easier	21.875%
There was no difference	0.0%
Combined was slightly easier	21.875%
Combined was easiest	43.75%

Table 3.5: Responses to which display method they felt was easiest to watch.

Response	Percent
Side-By-Side was preferred	25.0%
Side-By-Side was slightly preferred	25.0%
I have no preference	3.125%
Combined was slightly preferred	28.125%
Combined was preferred	18.75%

Table 3.6: Preferences for display method for a real search scenario.

This does not detract from the benefits of the fused display, but does show that it might be able to do even better with some improvements in the alignment.

The subjects reported an equal mix of opinions when asked which display method was easiest to find objects in (Table 3.4). This supports our results showing that subjects were able to perform equally on both display methods.

When looking at the results for which display was easier to watch (Table 3.5), we see a definite bias towards the Combined display, which supports the results on the secondary task.

The results for the final question, asking subjects which display they would prefer in a real Search and Rescue scenario (Table 3.6), look very similar to Table 3.4. There was not a significant preference for one display over the other. When looking at individual responses, the subjects' responses to this question mimicked their response to the second question (asking which display method was easiest to find objects in), which would indicate that subjects would like to use the display method they feel helps them best find people in a Search and Rescue scenario.

Chapter 4

Conclusions

This thesis presents a new method for calibrating a multi-spectral camera mount. This method allows users to calibrate the camera mount while they are also calibrating each camera, a step that is essential to many computer vision techniques. The calibration method works and produces aligned frames from IR and Visible cameras.

The image alignment allows us to fuse the imagery from these cameras together to display them to the user as a single display. Using this fused display, users were able to perform equally well as when using a display showing both the IR and Visible imagery separately. This greatly decreases the screen real-estate needed to display these video sources, making it simpler to use on small portable displays in wilderness areas. The fused display also decreases the cognitive load for the users, allowing them to better perform additional tasks simultaneously. This is very important in a search scenario, as the searchers need to be in constant communications with others to receive and pass information on as it is discovered.

This work has taken us one step closer to be able to use UAVs equipped with both IR and Visible cameras for Wilderness Search and Rescue. We can now work with both cameras on small display devices that will be used in Wilderness Search and Rescue. We have also increased the information the searchers can use without adding the increased cognitive load of watching two displays.

4.1 Future Work

Our work has shown great increase in the ability to do the search tasks given a fused display over a side-by-side display when snow is on the ground. There are other factors that would be interesting to look at with more time, such as fatigue, comparing performance and cognitive load against a simple Visible video. These tests were not performed but would give further strength to the validity of this type of display. The methods also need to be tested in different conditions such as during the summer or over lakes.

To be able to use these methods successfully during times such as the summer, when differences in the heat radiated from a person is not much greater than the surroundings, a better segmentation method for segmenting the IR may need to be employed. Work in the area of Infrared segmentation could be looked at to replace our simple thresholding. Also, some Visible spectrum segmentation methods might be extendable to Infrared imagery and could potentially be used.

Another factor that would be good to look at is different sensors. We had only a single IR and Visible camera at our disposal, each of which used adaptive gains and the Visible camera used automatic while balance control. Cameras that have more control over their internal gains and white-balance characteristics could greatly improve the quality of video returned from the cameras, adding to their effectiveness in a search.

While our work does a good job for alignment, there are times when there are still some inconsistencies in the alignment of the frames. In our system this is usually due to changing temporal synchronization problems. These problems come from the fact that the IR sensor has a slower response rate than the Visible sensor and that trying to capture data to disk from two video streams can overload the system. This could be overcome with on-the-fly registration adjustments. These adjustments may

be possible using a simple rigid body model or may require more robust models such as affine or perspective transformations.

The camera mount and camera calibration used in this work open up a number of new possibilities due to the frame alignment between the Visible and IR imagery. Many techniques that can be applied to Visible video can now be extended to the Infrared video, such as mosaicing, motion detection, and multi-frame noise removal.

Appendix A

User Study Instructions

Instructions

Please read before beginning the User Study.
You may refer back to them at any time during the study.

In the area of Wilderness Search and Rescue, aerial video has the potential to help out searchers in finding people and objects of interest. This assistance may possibly be amplified by adding additional types of cameras such as Infrared cameras that can detect heat.

This user study is aimed at evaluating methods to display a normal video (we refer to this as a visible video) and an infrared video simultaneously. To do this we will show you a number of videos. Some will have a Side-By-Side display while others will have a Combined display. The Side-By-Side display will show the visible video on the left and the infrared video on the right. The combined display shows the visible video with a hot-pink colored overlay that shows where hot objects are in the infrared video. This overlay automatically turns on and off at a regular interval to allow the user to see the original color of the object as well as see the heat information.

While watching these videos, we want you to look for two types of targets: (1) a person lying down and (2) red circles on the ground. Each video may have a different number of targets and some may not have any. These people or circles may show up in the video in a variety of ways. They may be detectible in only one of the types of video (visible or infrared) or they may be detectible in both. When you see one of these items we ask you to mark the location in the video. This marking is done by pressing the '**z**' key when you see a Person or pressing the '**x**' or '**/**' key (whichever is easiest to remember) when you see a Red Circle. When you press a key, confirmation text shows up at the bottom of the screen showing that the mark was made (If no text appears please remark the object). Example images as well as a practice section will be provided for you to look at in the user study software before data is collected. The training section allows you to go through all of the steps of the user study a few times before performing the actual study. After the completion of each video, you will be asked how many objects you feel you did not mark.

While watching the videos, you will also hear a number of tones played through the headphones. You will hear either a single tone repeated multiple times or two different tones repeated a number of times (before each video is played we will inform you of whether a single tone will be played or whether two different tones will be played). We ask that you count and remember how many times each type of tone is played. The tones will be easily discernable as one is a high pitch and the other is a low pitch. At the end of the video you will be asked to enter the number of times you heard each type of tone as well as if you felt that you missed any.

Before you start the user study, you will need to fill out the Consent Form, the Usability Test Compensation Form as well as the Preliminary Questionnaire. The Consent Form is a generic form used by our research group to inform you of possible hazards due to participating in the user study. The Usability Test Compensation Form is a form used to compensate you for your time and assistance. The Preliminary Questionnaire has a number of questions that we need answered to understand further if there is anything that could bias the results of the study. Please fill out the forms if you have not already done so and then begin the user study.

Thank you for your participation in this study.

Appendix B

User Study Pre-Training Questionnaire

Pre-Training Questionnaire

Please check only one choice per question.

1. Do you have any physical limitations that may possibly affect your performance in this user study (e.g. color-blindness, vision impairment, hearing impairment, impaired motor skills, etc.)?

O No
O Yes, Explain_____

2. How experienced do you feel that you are with using computers?

O Expert
O Average
O Novice

3. How experienced do you feel that you are with wilderness search and rescue tasks?

O Expert
O Average
O Novice

4. How experienced do you feel that you are with tasks involving searching for things on the ground from high up above in the air (aerial searching tasks)?

O Expert
O Average
O Novice

5. How familiar are you with the research related to this study?

O Never heard of any of it before this user study.
O I have heard about the research, but I have never seen any of the video display methods before.
O I know about the research, and I have seen the video display methods before.

6. How familiar are you with others' preferences of the display methods that you will be presented with in this study?

O I know nobody else's preferences.
O I know somebody else's preferences.
O I know a couple other people's preferences.
O I know many peoples' preferences.

Appendix C

User Study Follow-Up Questions

Follow-up Questions

Please check only one choice per question

1. In the combined video there are times when the overlaid heat information does not match the exact location of the objects in the visible video. Did this cause any difficulty in watching the video or understanding which object was hot?
   - O Yes
   - O No

2. Which display method do you feel was the easiest to find objects in?
   - O Side-By-Side was easiest
   - O Side-By-Side was slightly easier
   - O There was no difference
   - O Combined was slightly easier
   - O Combined was easiest

3. Which display method do you feel was easiest to watch overall?
   - O Side-By-Side was easiest
   - O Side-By-Side was slightly easier
   - O There was no difference
   - O Combined was slightly easier
   - O Combined was easiest

4. Which display method would be your preference in a real search situation?
   - O Side-By-Side was preferred
   - O Side-By-Side was slightly preferred
   - O I have no preference
   - O Combined was slightly preferred
   - O Combined was preferred

5. Any Comments/Concerns about this research (please explain)?

Bibliography

[1] KSL5 News, "Detective dies from helicopter crash injuries," November 2006. [Online]. Available: http://www.ksl.com/?sid=666341&nid=148

[2] D. Gibbins, P. Roberts, and L. Swierkowski, "A video geo-location and image enhancement tool for small unmanned air vehicles (UAVs)," in *Intelligent Sensors, Sensor Networks and Information Processing Conference*, 2004, pp. 469–473.

[3] M. Quigley, B. Barber, S. Griffiths, and M. A. Goodrich, "Towards real-world searching with fixed-wing mini-UAVs," in *IEEE International Conference on Robotics and Automation*, 2005, pp. 3028–3033.

[4] F. Rafi, S. M. Khan, K. Shafiq, and M. Shah, "Autonomous target following by unmanned aerial vehicles," in *SPIE Defense and Security Symposium*, Orlando, FL, 2006.

[5] M. Hansen, P. Anandan, K. Dana, G. van der Wal, and P. Burt, "Real-time scene stabilization and mosaic construction," in *IEEE Workshop on Applications of Computer Vision*, 5-7 Dec 1994, pp. 54–62.

[6] M. Tico, S. Alenius, and M. Vehvilainen, "Method of motion estimation for image stabilization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 14-19 May 2006, pp. II–II.

[7] D. L. Johansen, J. K. Hall, and C. N. Taylor, "Stabilization of video from miniature air vehicles," in *AIAA Conference on Guidance, Navigation, and Control*, Aug 2007.

[8] M. Irani, P. Anandan, and S. Hsu, "Mosaic based representations of video sequences and their applications," in *International Conference on Computer Vision*, vol. 00, Los Alamitos, CA, USA, 1995, p. 605.

[9] R. Szeliski, "Video mosaics for virtual environments," *IEEE Computer Graphics and Applications*, vol. 16, pp. 22–30, 1996.

[10] B. S. Morse, D. Gerhardt, C. Engh, M. Goodrich, N. Rasmussen, D. Thornton, and D. Eggett, "Application and evaluation of spatiotemporal enhancement of live aerial video using temporally local mosaics," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2008.

[11] B. Ready, C. Taylor, and R. Beard, "A Kalman-filter based method for creation of super-resolved mosaicks," in *IEEE International Conference on Robotics and Automation*, May 15-19, 2006, pp. 3417–3422.

[12] M. Kontitsis, K. Valavanis, and N. Tsourveloudis, "A UAV vision system for airborne surveillance," in *IEEE International Conference on Robotics and Automation*, 2004, pp. 77–83.

[13] L. Merino, F. Caballero, J. M. de Dios, and A. Ollero, "Cooperative fire detection using unmanned aerial vehicles," in *IEEE International Conference on Robotics and Automation*, 2005, pp. 1884–1889.

[14] J. Bradley and C. N. Taylor, "Particle filter based mosaicking for tracking forest fires," in *AIAA Conference on Guidance, Navigation, and Control*, Aug 2007.

[15] K. Hajebi and J. S. Zelek, "Dense surface from infrared stereo," in *Eighth IEEE Workshop on Applications of Computer Vision*. Los Alamitos, CA, USA: IEEE Computer Society, 2007, p. 21.

[16] S.-S. Lin, "Review: Extending visible band computer vision techniques to infrared band images," University of Pennsylvania Department of Computer and Information Science, Tech. Rep. MS-CIS-01-04, January 2001.

[17] J. Knight and I. Reid, "Self-calibration of a stereo rig in a planar scene by data combination," in *15th International Conference on Pattern Recognition*, vol. 1, 2000, pp. 411–414 vol.1.

[18] A. Zisserman, P. Beardsley, and I. Reid, "Metric calibration of a stereo rig," in *IEEE Workshop on Representation of Visual Scenes*, 24 Jun 1995, pp. 93–100.

[19] C. O'Conaire, N. O'Connor, E. Cooke, and A. Smeaton, "Comparison of fusion methods for thermo-visual surveillance tracking," in *9th International Conference on Information Fusion*, July 2006, pp. 1–7.

[20] F. Bunyak, K. Palaniappan, S. K. Nath, and G. Seetharaman., "Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking," *Journal of Multimedia*, vol. 2, pp. 20–33, August. 2007.

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

[22] J. Pluim, J. Maintz, and M. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[23] Y. Lin and G. Medioni, "Map-enhanced UAV image sequence registration and synchronization of multiple image sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–7.

[24] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[25] M. Irani and P. Anandan, "Robust multi-sensor image alignment," *Sixth International Conference on Computer Vision, 1998*, pp. 959–966, 4-7 Jan 1998.

[26] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[27] L. Zheng and R. Laganiere, "Registration of IR and EO video sequences based on frame difference," in *Canadian Conference on Computer and Robot Vision*, 2007, pp. 459–464.

[28] 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*, Aug 2003, pp. 228– 233.

[29] P. Rudol and P. Doherty, "Human body detection and geolocalization for UAV search and rescue missions using color and thermal imagery," in *IEEE Aerospace Conference*, March 2008, pp. 1–8.

[30] R. Raskar, A. Ilie, and J. Yu, "Image fusion for context enhancement and video surrealism," in *3rd International Symposium on Non-photorealistic Animation and Rendering*. New York, NY, USA: ACM Press, 2004, pp. 85–152.

[31] P. Burt and R. Kolczynski, "Enhanced image capture through fusion," in *Fourth International Conference on Computer Vision*, 11-14 May 1993, pp. 173–182.

[32] T. Wilson, S. Rogers, and M. Kabrisky, "Perceptual-based image fusion for hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 4, pp. 1007–1017, Jul 1997.

[33] M. A. Goodrich, B. S. Morse, D. Gerhardt, J. L. Cooper, M. Quigley, J. A. Adams, and C. Humphrey, "Supporting wilderness search and rescue using a camera-equipped mini UAV," *Journal of Field Robotics*, vol. 25, no. 1-2, pp. 89–110, 2008.

[34] P. van den Elsen, L. van 't Zelfde, and M. Viergever, "Near-automatic detection of arrow-shaped markers for CT/MRI fusion," in *11th IAPR International Conference on Pattern Recognition*, vol. I, 1992, pp. 755–759.

[35] J. Illingworth and J. Kittler, "A survey of the hough transform," *Computer Vision, Graphics, and Image Processing*, vol. 44, no. 1, pp. 87–116, 1988.

[36] A. P. Dempster, N. M. Laird, and D. B. Rdin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[37] J.-Y. Bouguet and P. Perona, "Camera calibration from points and lines in dual-space geometry," California Institute of Technology, Tech. Rep., 1997.

[38] ——, "3d photography on your desk," *Sixth International Conference on Computer Vision*, pp. 43–50, Jan 1998.

[39] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[40] E. Michaelsen, M. Kirchhof, and U. Stilla, "Sensor pose inference from airborne videos by decomposing homography estimates," in *International Society for Photogrammetry and Remote Sensing Congress*, July 2004.

[41] V. Bruce, M. A. Georgeson, and P. R. Green, *Visual Perception: Physiology, Psychology, and Ecology*. Psychology Press, August 2003.

[42] "Black Widow AV — Wireless Aerial Video Solutions". [Online]. Available: http://www.blackwidowav.com

[43] "FLIR Pathfind IR". [Online]. Available: http://www.corebyindigo.com

[44] "Procerus Technologies". [Online]. Available: http://www.procerusuav.com

[45] J.-Y. Bouguet, "Camera Calibration Toolbox for Matlab." [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[46] "OpenCV - Open Source Computer Vision Library." [Online]. Available: http://www.opencvlibrary.sourceforge.net