2006

THE SMART BOOKSHELF

Danny Sylvester Crasto
University of Kentucky, danny.crasto@gmail.com

ABSTRACT OF THESIS

THE SMART BOOKSHELF

The "smart bookshelf" serves as a test-bed to study environments that are intelligently augmented by projector-camera devices. The system utilizes a camera pair and a projector coupled with an RFID reader to monitor and maintain the state of a real world library shelf. Using a simple calibration scheme, the homography induced by the world plane in which book spines approximately lie is estimated. As books are added to the shelf, a foreground detection algorithm which takes into account the projected information yields new pixels in each view that are then verified using a planar parallax constraint across both cameras to yield the precise location of the book spine. The system allows users to query for the presence of a books through a user interface, highlighting the spines of present book using the known locations obtained through foreground detection and transforming image pixels to their corresponding points in the projector's frame via a derived homography. The system also can display the state of the bookshelf at any time in the past. Utilizing RFID tags increases robustness and usefulness of the application. Tags encode information about a book such as the title, author, etc, that can be used to query the system. It is used in conjunction with the visual system to infer the state of the shelf.

This work provides a novel foreground detection algorithm that works across views, using loose geometric constraints instead pixel color similarity to robustly isolate foreground pixels. The system also takes into account projected information which if not handled would be detrimental to the system. The intent of this work was to study the feasibility of an augmented reality system and use this application as a testbed to study the issues of building such a system.

KEYWORDS: Computer Vision, intelligent environments, human computer interaction, RFID

THE SMART BOOKSHELF

By

Danny Sylvester Crasto

Dr, Christopher Jaynes
Director of Dissertation

Dr. G. Wasilkowski
Director of Graduate Studies

April 11th 2006

THESIS

Danny Sylvester Crasto

The Graduate School

University of Kentucky

2006

THE SMART BOOKSHELF

THESIS

A dissertation submitted in partial fulfillment of the requirements of the degree of Master of Science in the College of Engineering at the University of Kentucky

By

Danny Sylvester Crasto

Lexington, Kentucky

Director: Dr, Christopher Jaynes, Department of Computer Science

Lexington, Kentucky

2006

ACKNOWLEDGMENTS

It has been a long arduous journey for me to reach this stage in my life. I came to the University of Kentucky as a undergraduate. It was a great opportunity to work on my Master's. I consider it fortunate affliction. It has been a learning experience that I have benefited from a lot.

I would like to thank Dr. Christopher Jaynes for his support and help with my research work that I am presenting here. He came up with this idea in his usual enthusiastic, casual, style, stating "Wouldn't it be cool if...". Without his support, I would not have been able to complete this. Thank you for the lab space, equipment, your valuable insight and genuine interest in my work.

Secondly, I would like to thank Dr. Amit Kale. With the amount of work he has put in with suggestions and discussions, I feel that he should be the co-author of this work. I would like to sincerely thank you for all your unselfish help. I don't think that I could have done this without you.

Last, but not least, I would like to thank my mother, who has been my cornerstone during my academic career. With her never questioning support, I would never would have made it. Thank you mom, I love you.

For Sylvester, Celine, Beulah and Moses Crasto

Contents

List of Figures

Chapter 1

Introduction

"The complexity for minimum component costs has increased at a rate of roughly a factor of two per year ... Certainly over the short term this rate can be expected to continue, if not to increase. Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years. That means by 1975, the number of components per integrated circuit for minimum cost will be 65,000. I believe that such a large circuit can be built on a single wafer."[10].

Moore's Law stated above, can be attributed as one of the reason why there is an increased growth of research and development in computer vision. Essentially, computing power has caught up to the processing demands of the field. The exponential growth of computer "complexity" allows computers to be powerful and cost effective enough to process large amounts of data at real time rates. Advances in cameras have allowed computer vision applications to see the real world environment around them in real-time. Coupled with availability of low costs projectors, vision applications are able to interact with human actors in their native environment without any perceived delay.

Recent research in projector-camera systems has overcome many obstacles to deploy and use intelligent displays for a wide range of applications. In parallel with these developments, projector costs continue to decline with corresponding increase in resolution, brightness and contrast ratio. In light of this trend, our work explores how projector-camera systems can be used to continuously monitor and appropriately augment a changing scene. To better explore the unique capabilities that these systems can offer to intelligent and ubiquitous computing environments we introduce a simple application, The Smart Bookshelf.

The system was designed so that it could be useful for individuals who want to find out if books were present or absent on a crowded bookshelf, quickly. For this to occur, the system would be required to know the state of the shelf and also the precise locations of books present. Traditionally users would look up a book code and search the sorted shelf for the corresponding book. Using this analogy, we also wanted to allow the user to be able to query for books but relax the sorted constraint as books are not always sorted on a library shelf. The system needs to provide quick visual feedback to indicate successful and unsuccessful queries. With these usability concerns in mind, the Smart bookshelf supports the following tasks.

- Detect addition, removal and reinserted of books by the user.
- Allows users to query for books by title or author, highlighting books that are currently present and indicating absence of books if they are not present.
- Re-project the contents of the shelf at a specified time.

Implementation of these three straightforward tasks would allow users to more effectively locate books, discover when they have been removed, and track down titles in a large library.

The system utilizes a pair of cameras and a mounted projector to achieve the tasks stated above. This work demonstrates that the presence of a camera can be utilized to accurately monitor scene state (geometry, presence/absence of objects, etc.) in addition to the traditional utilization of a camera in order to calibrate and correctly render information via a projector. The state of the shelf is also inferred by using RFID tags which encodes information about books such as the author, title, etc that can be used to query for books. Working in tandem with the visual system, the tags only provide information about presence or absence of books that are within 10 feet of the tag reader. The visual system is triggered by the state change and accurately determines the location of books in the shelf. A calibrated projector is utilized to augment the scene to allow the user to easily find books in the shelf. It maintains a database of books that are currently present or were present in the shelf along with their locations. The state information along with its corresponding precise location, allows the user to query the system and indicate whether a book is present by highlighting the book or indicating to the user that the book is no longer present using the projector.

The idea of augmented environments utilizing projectors is not a new concept. An early system that combined projected information in a real world environment was the "digital desk" described by Wellner and MacKay [16]. The idea was to augment a regular working desk by projecting electronic information as well as to monitor paper documents to facilitate a seamless integration of information. It looks to fuse the world of virtual and the real making human computer interaction (HCI) ubiquitous. Initial immersive HCI was done utilizing head mounted displays. The human actor would be transported into a synthetic world with this device which was cumbersome, intrusive and not natural to the user. Current vision applications strive to be part of the real world and interact with it transparently as it has become feasible from a monetary and efficiency standpoint. The medium for the interaction is the real world environment that the user is in and means is achieved by utilizing projector and cameras. The imagery being displayed on real-world surfaces such as walls, furniture, etc, can also be viewed by more than one person, allowing for multi-user user interaction. Such a multi-user application is described in [15] which utilizes multiple projectors and interacts will multiple actors rending objects on a curved rear-projected surface

A significant amount of progress has been made in projected displays that utilize a camera to monitor projected imagery as well as the scene into which it is being projected[5, 11] . These research efforts have addressed many significant problems related to projector-based augmented reality including correct color production [9], geometric warping and blending [4], multi-projector cooperation [4, 7], resolution requirements, and blending of projected imagery with underlying surface characteristics [11]. Much of this work is being conducted within the computer vision and graphics communities and as progress is made, projector-based augmentation is becoming more feasible.

[12], Pinhanez et al. propose the use of steerable projector-camera systems called Everywhere Displays that allow people to interact with projections which included games, arrows that provided direction and their other applications that include a ubiquitous product finder for retail environments. Raskar et al.[4] "The Office of the Future"propose the use real-time computer vision techniques to dynamically extract per-pixel depth and reflectance information for the visible surfaces in the office including walls, furniture, objects, and people, and then to either project images on the surfaces, render images of the surfaces, or interpret changes in the surfaces. More recently, Crowley [3], proposed a steer-able camera-projector system with two degrees of freedom which can be used to detect and track a planar surfaces for display.

It is an exciting time to be working in the projector-camera domain of computer vision as previous research, availability of computing power and low cost projectors have opened up avenues to unexplored areas which have not being ventured into, allowing for real-time and transparent HCI. There are already a diverse number of fields in which augmented reality projector-camera systems have being utilized. In the medical arena systems such as the REVEAL project[13], would allow doctors to use the generated imagery from laparoscopy procedures as an instructive tool and also be able to delocalize the surgeon from the procedure, allowing for the possibility of remote surgeries. Larger scale interactive, immersive, environments are now possible. Systems such as the "metaverse"[7] where tracked users can interact with their synthetic environment have already been built. Such a system could be used as a training tool by the military to put soldiers in dangerous environments to teach them to survive. The possibilities seem boundless.

1.1 Contributions

The contributions of this work is primarily an exploration of the system, algorithmic, and human-computer interaction issues in realizing a projector-camera system in the real world. This work does not claim that the vision components are revolutionary. They are made to operate in somewhat constrained conditions. However, we introduced techniques that take advantage of these conditions and take into account the changing projected illumination to increase the robustness of the vision algorithms. In particular the contribution of the thesis are:

1. A working testbed for projector-camera systems in a real-world application to support further research.
2. A robust book-matching algorithm that utilizes geometric constraints to detect changes in the area of interest which takes into account possible augmented information which is projected into the scene.
3. A cooperative multi-sensor approach utilizing RFID with vision to maintain bookshelf state as well as an accurate appearance model.

This work is divided as follows. Section 2 discusses in detail the methodology adopted. Section 3 discusses the main results of the work. Lastly, Section 4 is the conclusion, summarizing the main findings and recommendations for future work.

Chapter 2

Proposed Methodology

The goal of this work is to create an application that monitors books that are added, removed and reinserted into a bookshelf. Using a projector, the system should be able to augment the scene in response to user query's for books that may or may not be present in the shelf at the current time. It should also have the ability to display the state of the shelf at a given time. A two phase calibration step described in (2.3) is performed prior to system use that will support the system to achieve these requirements. Once setup, the system will run infinitely, monitoring the scene for change and allow users to interact with it.

2.1 System Setup

The flow chart above portrays the working system that was created. Although we describe details for an experimental setup, many of these details could be altered for an applications needs. For our system, there are two firewire cameras mounted on tripods placed approximately three feet apart. The tripods are at the same relative height. The cameras are placed about five feet away from the shelf. A projector is mounted on a tripod and placed in between the cameras. The cameras and the projectors are oriented so that their frustums provide coverage of the shelf to be monitored. There are two machines used in the system. The main server, is the machine that the cameras are attached to. The slave machine is attached to the projector. The server receives RFID messages from a reader over the network and maintains the state of the shelf in a local database. It also serves as a terminal that allows users to query the system. The server sends signals over the network to a slave machine to highlight respective areas in response to positive user queries which correspond to books in the shelf or notify the user of unsuccessful queries.

The flow chart shown above depicts how the state of the shelf is updated and maintained by the system. The initial state change is determined by processing RFID packets that are sent by the reader to the system over the network. It compares the

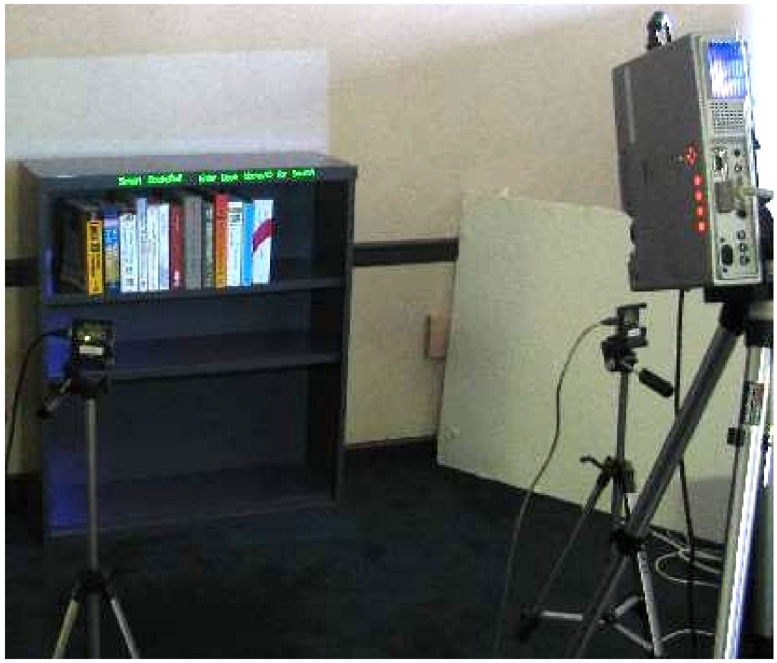


Figure 2.1: System setup with two cameras and a projector

previous known list to the RFID generated list to determine the state change. If there is no difference in the lists, nothing is done. However if there is a discrepancy, the state change can easily be deduced. If a book is removed the foreground is updated and the book is placed in the absent list. This is described in detail in (2.7). If a book is added or reintroduced into the system, section (2.6) describes how the state is updated.

The equipment used to build the system was

- Main server: Xeon 2.4 GHz, 1 GB RAM, Ethernet, 40GB harddrive. On-board IEEE 1394 6pin.
- Slave machine: 2.4 GHz Pentium 4, 1 GB RAM, Ethernet, 40GB harddrive.
- Two Point Grey Firewire Cameras, transmitting 640x480 YUV 4:1:1 images at 30 fps mounted on tripods
- Epson Project 1280x1024 Native resolution with VGA input.
- Alien Technologies RFID Reader (ALR-9780).
- Alien Technologies RFID Adhesive Tags (ALL-9354-02).

To have a usable system up and running the number of events that occurred during a given instance was restricted. Also, the allowed orientation of books placed

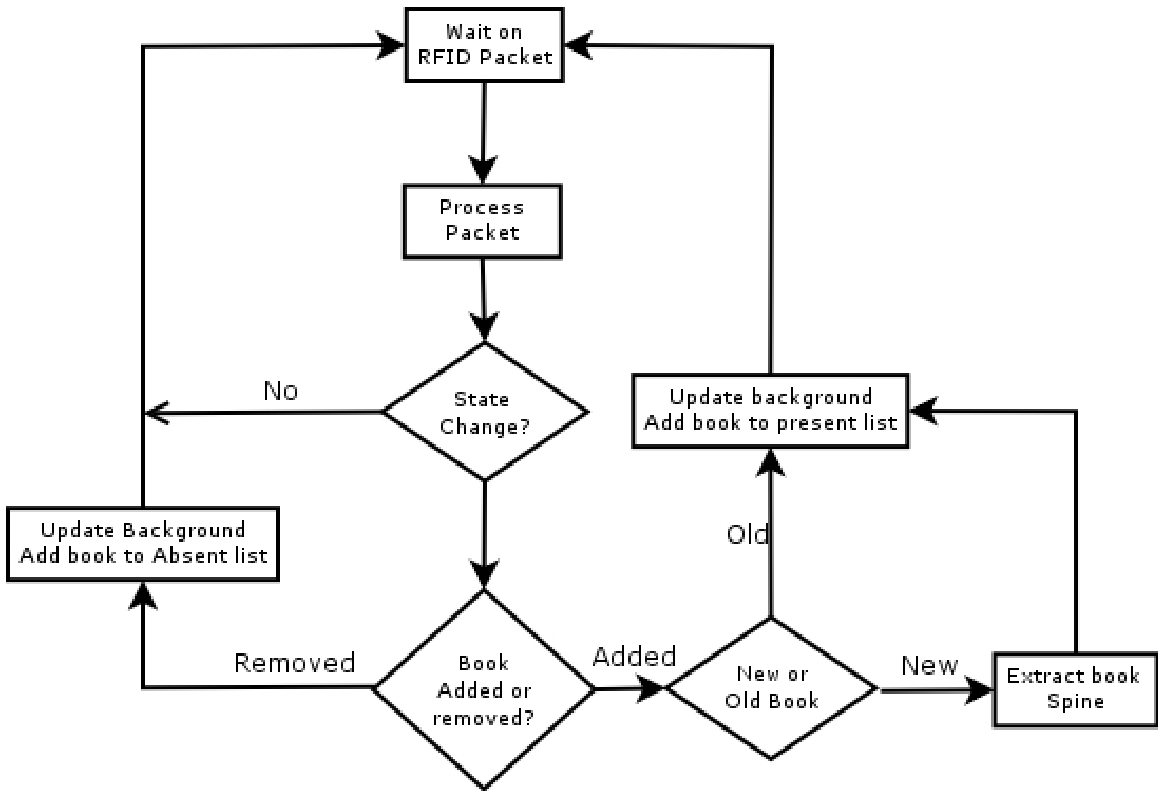


Figure 2.2: System flow chart

on the shelf was restricted to the spines of the books facing outwards and upgright. Explicitly, the assumptions were:

- One event occurs in the given time interval. The event is either a book addition, removal or reinsertion. The length of the time interval is the duration between receipt of the RFID packets from the network.
- Books on the shelf are placed upright with the spines facing outwards.

The main focus of this work is to provide this testbed and having a single event occurrence simplifies our vision detection tasks but leaves the door open for future work of detecting multiple events with a combination of RFID and vision techniques. The RFID system controls the length of an event interval which was set to 20 seconds, which is a more than reasonable time for an event to occur. The second assumption generally conforms to the use of a real-world bookshelf. Having the spines of the books facing outward is important as that is how users would discriminate between books on a shelf. We use this assumption to geometrically extract the spines of the books as described in section 2.2.

In the systems infancy, it was understood that to build a robust system, the spine pixels of books had to be isolated. Isolating these areas is not a trivial task due to the unconstrained positions of the cameras which will include non-spine portions namely the sides and tops of the books. These non-spine regions are a significant problem. They had to be discriminated against for the following reasons.

- Firstly, the non-spine pixels may be in shadow and perspectively distorted with respect to the cameras. This can lead to unstable color models which are used for matching books and cause false positives.

- Secondly, we wish to augment the shelf with the projector. These non-spine portions maybe occluded by neighboring books and if included in the model would result in erroneous augmentation which now includes a region larger that the actual book. To illustrate the point, consider a book which is added to an empty bookshelf and the resulting foreground detection via background differencing (shown in Figure 2.4) is obtained. When a second book is added to the bookshelf, the side of the first book is no longer visible due to occlusion by the second book. Subsequent augmentation of the scene results in a highlighted region significantly larger than the visible region of the book.

- Lastly, the spines of the books typically provide a unique characterization in terms of text, color and intensity patterns which is can be utilized for recognition purposes.

This was the main challenge of the project and a significant amount of vision processing is dedicated to isolating the spines of books which along with other information such as RFID id, title, author, etc, is associated with a book in our system.

With a single camera, it was hard to disambiguate the observed sides from the spines as they are closely similar in appearance, and would vary with book placement on the shelf relative to the location of the camera. This can be seen in Figure . We sort out a more robust and elegant solution that involved a multi-camera system to isolate the spines of the books. From common experience we know that books are stacked on bookshelves with their spines facing the user as shown in Figure 2.3. We can see that their spines lie in the same approximate world plane π that passes through the front of the bookshelf. So adding a second camera would allow us to use a homography which is induced by this (approximate) plane to segment spine regions from regions outside of the plane. Using a combination of background subtraction and a technique referred to as "homography filtering", spine regions of the books can be reliably identified. In order to exploit geometric constraints imposed by the assumption that book spines lie in a plane, we must first estimate the geometric relationship between the cameras. It consists of discovering the homographies that map pixels from one device to another that correspond to this plane. The second calibration phase is obtaining a set of color transfer functions for each of the devices.

Figure 2.3: Two views of the bookshelf seen from monitoring cameras. Color characteristics of the book spines are used for recognition. Multiple views of the plane containing the spines help segment book spines as they are added and removed from the scene.

(a) (b)

Figure 2.4: (a) and (b) show a single book added to the shelf and the corresponding foreground pixels detected by a background subtraction algorithm. Notice that sides of the book are included in the detected foreground.

These functions map projected colors to observed colors in each of the cameras. These functions are used during image processing to obtain the background model of the scene. It is essential that the algorithms take into account of projected information which if not handled would lead to inaccurate background models as described later. Figure 3.1(a) shows a picture of our physical setup as well as a schematic view of the situation. The calibration of the system is discussed in section 2.3

Noting that the spines of the books are more or less coplanar, this assumption infers a very strong constraint exists for the corresponding spine points in the left and right images of the stereo pair [6]. Specifically, if $\tilde{\mathbf{m}}_L$ and $\tilde{\mathbf{m}}_R$ denote homogeneous coordinates of a spine point in the left and right views respectively then

$$\tilde{\mathbf{m}}_L = H_{LR}\tilde{\mathbf{m}}_R \tag{2.1}$$

where H_{LR} is a collineation which is a function of the location of the spine plane and the camera internal and relative external parameters.

The initial direct approach that we resorted to was to consider the foreground images as in Figure 2.3 obtained in each of the cameras and then performing cross-validation for potential correspondences using Equation (2.1) to extract small image patches and measuring local color similarity. This technique has been successfully used in other contexts [2], but requires color calibration of the different cameras because of potentially dramatically different observed colors in each of the cameras. In addition, this approach is sensitive to specular highlights and other artifacts that lead to nonlinear (and different) responses in each view. For this application, we employed commodity cameras and do not want to burden the user with sophisticated camera calibration.

2.2 Homography Filtering

We employ a technique that utilizes a geometric constraint where the spines of books in the shelf lie on the same approximate world plane which governs the planar parallax across the two cameras and is used as constraint on segmenting book spines. Due to the domain, we can assume that the objects under consideration are rectangular. Hence binarized left and right foreground images when transformed to between views via the relevant homography will overlap in the regions of the spine. Conversely, non-spine regions conform to planes that are orthogonal (sides, top and bottom) to the spine plane. Thus under the collineation for the spine plane, non-spine pixels will not coincide over the images. This "homography filtering" technique allows us to utilize a more robust geometric constraint to isolate spine pixels rather than resort to less reliable image processing techniques such as image morphology. More importantly, this approach eliminates the need for color calibration across the cameras as we are only checking for presence and absence of foreground detected pixels in these binary images.

However, the constraint that all the book spines to lie in the same plane may seem unreasonable at first glance, for regular usage of a bookshelf as books can be placed at arbitrary depths in the shelf. Here we explore the implications of this constraint and its tolerance to the deviations from the assumed world plane. To consider the fact that book spines may defer in depth with respect to the cameras, consider the homographies H and $H + \Delta H$ and induced by the planes $n^T X + d = 0$ and a plane parallel to it but at a slightly different depth $n^T X + (d + \Delta d) = 0$ respectively. Following [6] we can show that

$$\Delta H = K'(tn^T)K^{-1}\frac{\Delta d}{d} \tag{2.2}$$

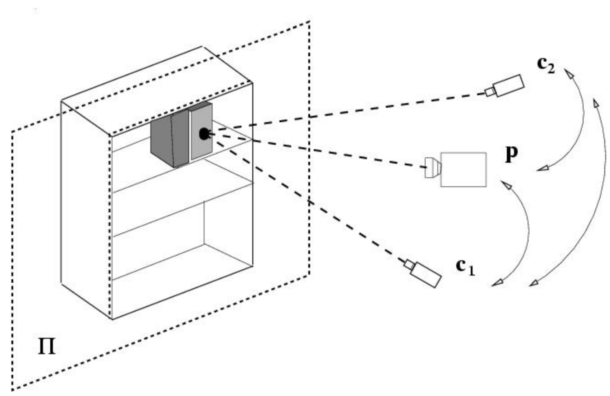


Figure 2.5: Using two cameras to effectively isolate the spine plane of the books using homography filtering

where K, K' represent the internal parameters of the cameras n, d represent the world plane coordinates of the assumed spine plane Π, t represents the translation between the cameras. As $\frac{\Delta d}{d} \rightarrow 0$ $\Delta H \rightarrow 0$. We evaluated this hypothesis in our experiments and it turned out that the degradation of performance in fact was not significant.

2.3 System Calibration

It is necessary to perform a two stage calibration of the the system. First, the pixel mapping between the left to the right camera and to the projector relative to the spine plane must be recovered. Secondly, a set of color transfer functions that describe the mapping of the color gamuts of the projector and cameras must be carried out. Geometric calibration is used directly during the scene augmentation phase to correctly illuminate books and other parts of the scene (detected in the camera frame) using pixels in the frame-buffer of the projector. Color calibration is required to improve the robustness of image processing routines in the presence of projected information (see Section 2.4).

In order to obtain the homography between the two cameras, several corresponding points on the spines of the books in the two cameras were collected and the direction linear transform method [6] was used to obtain the homographies H_{RL} and H_{LR}. In order to compute the homography between the projector and the cameras, a white board is placed against the bookshelf. Random dots are projected onto the screen and their positions in the camera frame are recorded. Using these point correspondences the homographies H_{LP} and H_{RP} are computed as before. This approach is in common use for projector calibration and can yield an accurate estimate of the geometric mapping between the two views.

A second phase of calibration estimates the color transfer function between the projector and the camera a technique described in [8] is used. The technique provides

Figure 2.6: Show the point correspondences for the left and right views respectively which is used for obtaining the collineation between cameras

a rough estimate of relating observed and projected colors. The three colors channels (Red, Blue and Green) are assumed independent and are calibrated separately for each camera. For each color channel, increasing intensities are projected and recorded. The other channels are set to zero. For each observed color the average intensity is calculated over ten trials and is used as the observed color for the given projected color. Using Levenberg-Marquardt nonlinear optimization technique, a fit is determined to the measured data points for each color channel in the form

$$f_c(x) = \frac{a}{1 + e^{-\alpha(x-b)}} + k \tag{2.3}$$

Utilizing this color transfer function we can determine the color of the projectors frame-buffer as they would appear in the camera. Far more sophisticated characterization of the color relationship between cameras and projectors is the subject of on going research [9]. However, for our purposes a rough estimate is all that is required for the system to detect changes on a expected background that is polluted with projected images 2.4.

2.4 Foreground Detection in the Presence of Projected Information

As augmentation takes place in the environment, the algorithms we designed to take into account projected information. Foreground detection was not accurate when there was projected information present in the scene. These projections would be included in the background subtraction result causing the foreground detection algorithm to fail. The initial approach directly wrapped the projector's frame buffer into the corresponding camera view and added the ON pixels to the camera image. This

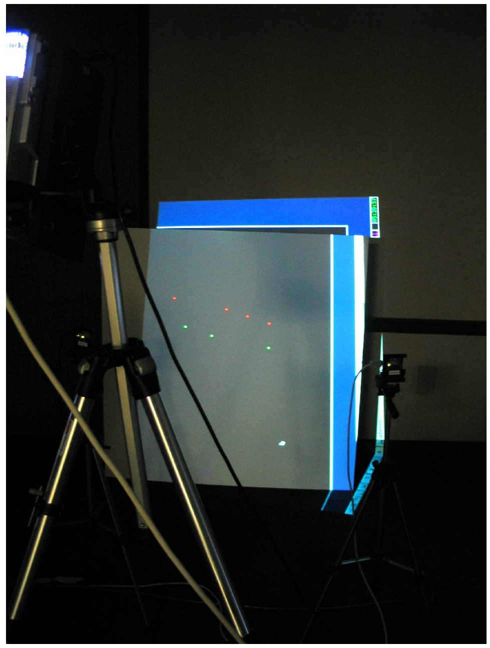


Figure 2.7: Homography estimation utilizing a board that is an estimate for the spine plane.

did not work well due to the differing color gamuts of the the camera and projector. Color transfer functions from projector to cameras were derived to compensate for the color difference. Now, in addition to the location warp, colors were mapped to the color-space of the corresponding cameras. This led to more accurate foreground detection. In practice the projected information is included in the background model prior the foreground detection process. The contents of the projectors frame buffer are geometrically warped to the corresponding locations in each of the cameras frames utilizing the derived homographies. Each projector pixel is is mapped to the camera position via $x_c = H^1_{pc} x_p$. The color value of the projector pixel, x_p is added to the color values that is already stored at camera pixel value x_c. The resultant is mapped through a sigmoidal function that is recovered during the color calibration phase. The new synthetic background image that is the generated takes into account any projected information. Figure 2.4 depicts the foreground detection process in the presence of changing augmented information. The scene was captured as a book as being added to the shelf. At the same time, the results of a previous book search still reside in the frame-buffer of the projector and highlight the scene. By combining the geometrically corrected and color mapped values in the frame-buffer into the space of a monitoring camera shown in Figure 2.4, this information can be used to construct a synthetic background image by overlaying information with the current background model

The above foreground detection subsystem is a key component in the overall algorithm and is used to pinpoint the presence of new objects as well as objects that have been removed from or moved on the shelves. By taking into account projected information in the background model, the system can continuously monitor the scene while the augmentation system is in operation. The following sections describe the main events that the system can handle and how it is done.

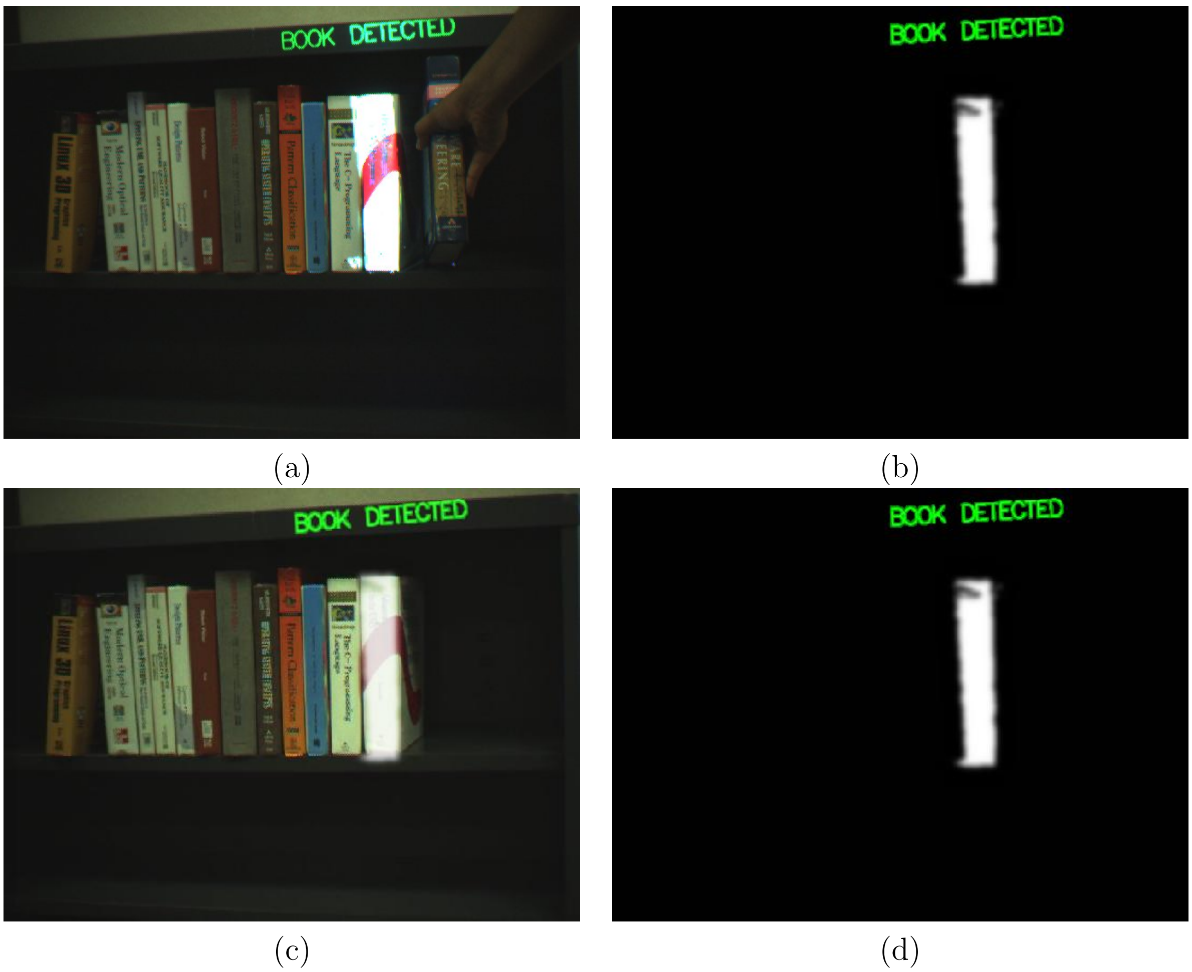


Figure 2.8: Foreground segmentation in the presence of augmented imagery. (a) Scene containing overlaid projected information as well as "real" change due to a user placing books. (b) Frame-buffer contents of the projector after they have been warped and color corrected for the left camera. (c) Synthetic expected image constructed after combining warped and corrected frame-buffer contents with background model. (d) Detected differences.

Figure 2.9: The images shows the incorrect computed bounding box indicating the changed region after a multi-event occurrence of books falling.

2.5 Augmenting vision with RFID

Adopting a vision only approach has it shortcomings. For example, when book spines had similar histograms, it would be hard to disambiguate one from the other. To detect if a change occurred, the system runs a foreground detection algorithm and finds the largest connected component. The area is thresholded to account for noise. This leads to the possibility that a book would slip through the threshold. Subsequently, if a large visual change occurred with no change in state, such as books falling, the system would not be able to determine what happened based on visual cues alone where the largest connected component would not correspond to a known book spine. This is illustrated with the Figures 2.9. These deficiencies lead to an unknown state which would be detrimental to the usability of the system.

In view of the issues described above and given the constrained nature of the domain, other sensors maybe feasible to help us over come these issues. An RFID sensor was last component that was added to the system. RFIDs is an emerging technology that provides identification of inanimate objects through the use of adhesive tags and a RFID reader. Using an omni-directional antenna the RFID reader maintains a field of radio waves at a frequency of 900 MHz and a radius of 10 feet. Tags are either active or passive. Active tags have a power source and an antenna that output the tag's ID at a fixed time interval. These tags have longer range but are much larger than their passive counterparts. Passive tags are labels that have imprinted antennas. They only advertise their ID when they are within the readers field by absorbing some RF energy. These tags have a longer shelf life and are smaller but they have a shorter range. The latter was used in this system.

Addition of this component was beneficial to the system. as it could now definitively obtain the state of the shelf several times a second. This could be used to handle the ambiguous situations that arose when a books looked similar or books that were too small were introduced into the shelf. Also, explicitly knowing the state of the shelf, the system has the ability to tackle the problem of a common occurrence of books in shelves, books falling over. These efforts were saved for future work of the project and is described in Section 4.

Added benefits were also realized with the addition of the new component that further strengthen the reasons for its inclusion. Computational time could be saved as the system would be notified of a state changes via the RFID system. Previously, the system would run a foreground detection algorithm every minute to determine if a change of state occurred. Additionally, if a user happened to be in the scene when the foreground detection algorithm was running, the system would have to discard that reading and wait for the next. With the RFID component, addition, removal and replacement events are determined robustly, efficiently and with finer granularity. Since each tag is unique, using a stand alone program, the tags of books could be rescanned and the information associated with each one, such as the title, author, ISBN, etc, could be input before hand. This would alleviate the responsibility of the user to input this information when a new book is added to the system making the it more useable.

However, the RFID system has poor spatial accuracy and therefore cannot be used on its own. It is only able to query tags that are within its field. The size of the field varies with the types of tags. For the passive tags that are utilized for this system, the detection field is 5 to 7 feet in radius. It can accurately determine which tags, and therefore books, are in the field. The visual component is needed to give the system spatial accuracy that is needed to allow users to query the system. Thus using the RFID system in tandem with the visual system makes for a more robust, efficient and user-friendly system.

Before incooperating the RFID system, the vision system was the only method used to determine whether a book was removed or replaced. However with the introduction of the new subsystem, this information was already provided and this determination could be made conclusively. All that needed to be done was update the current foreground image to reflect the appropriate state. It is still assumed that a single change occurs over a fix time interval. Before this new subsystem, the vision system was used to determine if a book was added, removed or replaced. The algorithms described below explains what was done prior to the addition of the RFID component.

2.6 Adding books to the bookshelf

For simplicity of explanation let us consider that the bookshelf is empty initially and a user is adding books to the shelf. Let ESL and ESR denote the images of the empty shelf in the left and right cameras respectively. After the first book is added, the system would be notified of the new addition and background subtraction based on thresholding the norm of the difference in the RGB values of the corresponding foreground and background pixels is used to obtain a foreground image as shown in Figure 2.4(b). If projected information is present, foreground images are generated using the process described in Section 2.4 considering the frame buffer contents. Let FL_1 and FR_1 denote the foreground images for the left and right camera respectively, H_{LR} and H_{RL} denote the collineations transforming the right camera pixels to the left one and vice-versa respectively. In order to extract the spine in the left camera we do the following: The ON pixels in the left camera are transformed to the right camera via H_{RL}. The source pixel in FL_1 is turned off unless the corresponding pixel in FR_1 is ON to get SPL_1. To summarize:

$$SPL_1(\tilde{\mathbf{m}}_L) = \begin{cases} 1 & \text{if} \quad FL_1(\tilde{\mathbf{m}}_L) = 1 \text{ and} \\ & \quad\quad FR_1(H_{RL}\tilde{\mathbf{m}}_L) = 1 \\ 0 & \quad\quad \text{otherwise} \end{cases} \tag{2.4}$$

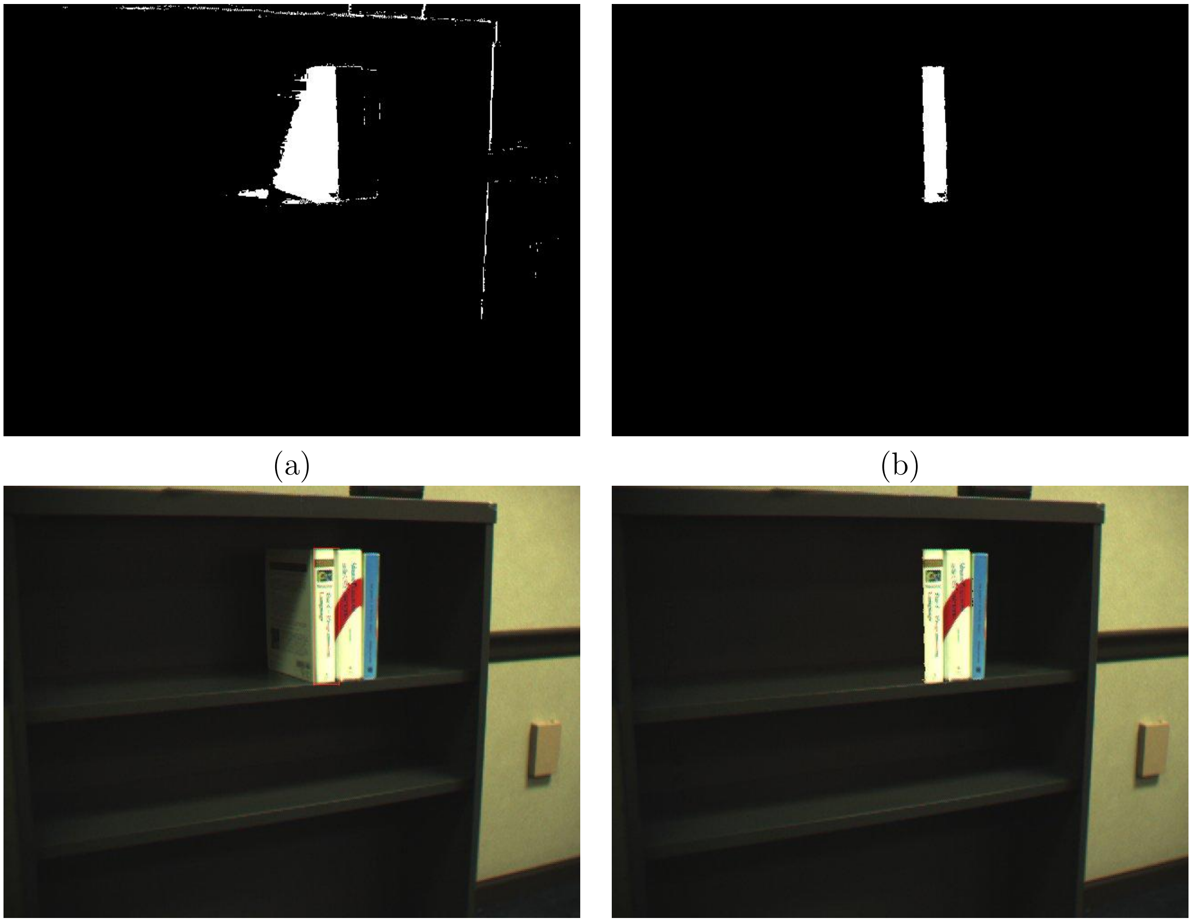
The largest bounding box $BL_1 = \{xL_1, yL_1, sxL_1, syL_1\}$ containing the resulting "homography filtered" image is then computed, where (xL_1, yL_1) denote the coordinates of the top left corner of the bounding box and sxL_1 and syL_1 denote the width and height of the bounding box. The color histogram HL_1 of the foreground image FL_1 in the region marked by BL_1 is computed and stored. Similar information for the right camera is computed as well. Also the time stamp of the book addition event is recorded. The user is then prompted to enter the title and author of the new book. The book along with the associated bounding box, time of capture and the color histogram is entered into the database.

If a second book is added to the shelf, the spine information for the first book must be taken into account so that the sides of the first book are not included in the background when detecting the new foreground. To do this base background ESL is augmented by the pixels inside bounding box BL_1 superimposed on the new foreground

$$EL_1(i,j) = \begin{cases} FL_1(i,j) & \text{if} \quad (i,j) \in BL_1 \\ ESL(i,j) & \quad\quad \text{otherwise} \end{cases} \tag{2.5}$$

The augmented EL_1 is an element of the current state $S(1)$ of the bookshelf. The state $S(t)$ of the bookshelf also includes other information such as identities of books currently present in the shelf, their locations and bounding boxes in both cameras.

After the user adds a new book, the updated background EL_1 is used to obtain the foreground images $F_L(2)$. The planar parallax technique, described before, is used to obtain the bounding box for the second book and the second book along with the associated bounding box, time of capture and the color histogram is entered into the



(a) (b) (c) (d)

Figure 2.10: Images of a bookshelf taken when a book is added. (a) Foreground image detected using the previous state of the bookshelf . (b) Foreground image after eliminating non-spine pixels using (2.4) (c) Bounding box BL_j computed from (c) superimposed on the foreground image for computation of the histogram HL_j for the book j. (d) Updated state of the bookshelf as described by (2.6).

database. The new augmented background EL_2 is computed as

$$EL_2(i,j) = \begin{cases} FL_2(i,j) & \text{if} \quad (i,j) \in BL_2 \\ FL_2(i,j) & \text{if} \quad (i,j) \in BL_1 \\ ESL(i,j) & \quad\text{otherwise} \end{cases} \tag{2.6}$$

The state of the bookshelf is updated to include EL_2 as also information about the second book. In this way by knowing the state of the bookshelf at any time t, new books can be added to the database. Figure 2.10 shows the steps in extracting the spine region of a newly added book.

2.7 Tracking book removal and replacement

When a book is removed from or replaced back into the bookshelf, the last known state of the bookshelf $EL_t \in S(t)$ is used to obtain the foreground FL_{t+1} similar to the way it was done during building of the bookshelf. For the case when a book is removed the image of the region where it was removed from will contain parts of the sides of the other books or the background.

Once again the planar parallax technique described by (2.4) can be used to extract the "spine"portion of the foreground with the implicit understanding that it corresponds to the region where a book used to be if the change event was a "Book Remove". A bounding box $BL = (x, y, sx, sy)$ is computed for the resulting region and a color histogram $HL(t+1)$ is computed for the corresponding foreground image. If the change event was a "Book Replace" event, then $HL(t + 1)$ is expected to be close to HL_i if book i was the book that was replaced. On the other hand if the change event was a "Book Remove" event then $HL(t + 1)$ would correspond to a foreground region which includes the sides of the neighboring books and $HL(t + 1)$ wont be close to any of $\{HL_1, \cdots, HL_N\}$. Specifically we compute

$$d^* = min_{i=1,\cdots,N} d(HL, HL_i) \tag{2.7}$$

If $d^* < T_{hist}$ then $i^* = argmin_i d^*$ is declared to be the book added. If $d^* > T_{hist}$, then the event is identified as a book removal event. In the absence of noise if the book m located at (x_j, y_j) was removed, then $x = x_j$ and $y = y_j$. To robustly detect which book was removed, we find the books $p \sim x_i$ and $q \sim x_{i+1}$ such that $x_i < x < x_{i+1}$. If book p was the one that was removed, then the region given by (x_i, y_i, s_x, s_y) in EL_t and in FL_{t+1} will be significantly different when compared using an appropriate histogram distance, while the region given by $(x_{i+1}, y_{i+1}, s_x, s_y)$ in EL_t and FL_{t+1} would be nearly identical. If we denote the histogram distance in the first case by d_i and the second case as d_{i+1}, then the book removed can be declared as

$$r \sim \text{argmax}_{i,i+1}(d_i, d_{i+1}) \tag{2.8}$$

where any appropriate histogram distance measure such as Bhattacharya distance or χ-squared distance [1] can used for computing the histogram distance.

In both cases the state of the bookshelf $S(t)$ is updated by adding appropriate information as regards replacement or removal.

2.8 Augmenting the smart bookshelf

During the system setup, the user is required to select two areas of the scene. The user is shown a camera image of the bookshelf and is required to select the four points

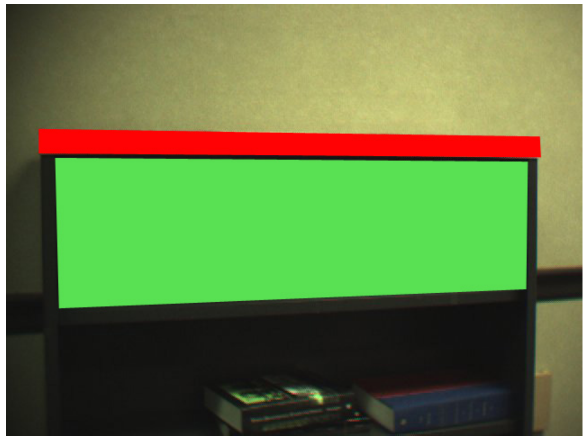


Figure 2.11: User specified regions of interest. The Red area is where system text messages are displayed. The Green area isolates where the books will be placed.

that bound each region. The first area is designated as the region where the system communicates with the user. This area is used to display a welcome message and responses to queries indicating presence or absence of books. The second area is the region where the books are placed, giving the user flexibility of specifying a particular area and also reducing computational costs of our vision algorithms by decreasing the working area. The mappings from cameras to projectors are known, allowing us to transform the points between coordinate frames. The Figure 2.11 shows the two user specified regions.

Information about the past states of the bookshelf as well as information regarding a specific book can be made available easily to the user using the projector. The most common task that the system should perform is to facilitate access to a book requested by the user. For instance, suppose a particular book i is requested by the user, then the system checks the current state of the bookshelf $S(t)$ to check if the book i is present. If it is present, knowing the homography induced by the spine plane between the projector and the left camera H_{LP}, the bounding box BL_i corresponding to the ith book is prewarped and pixels contained inside it are turned ON

$$P(i,j) = \begin{cases} 1 & \text{if} \quad (i,j) \in H_{PL}BL_i \\ 0 & \quad\ \ \text{otherwise} \end{cases} \tag{2.9}$$

If the book is not present then its "checked out" is displayed. Apart from highlighting the book requested by the user (Figures 3.1 (c),(d),(e)) several other visualization tasks regarding the library can also be performed. For instance, the user may request the state of the bookshelf or a particular book at an arbitrary time t (Figure 3.1 (f)). Furthermore auxiliary information about a book may also be projected into the scene. The projected information is taken into account during the subsequent image processing tasks as described in Section 2.4.

Chapter 3

Experimental Results

The smart bookshelf was implemented in our lab using a pair of tripod mounted cameras and a single projector. The baseline between the cameras was approximately about one meter. The projector was mounted on a tripod and placed in between the two cameras taking care not appear in the camera view as shown Figure 2.2. The projector and camera were aligned by hand to ensure that their frustums overlapped.

We first estimated the homographies induced by the spine plane between the cameras and projectors as discussed in Section2.3 and the color transfer functions between each of the camera and the projector as described in Section2.4. Users were told to add books to the bookshelf at different times as a part of the library building stage described in Section 2.6 and to query for the presence of a book and to remove or replace books back into the shelf. The system updates the states of the bookshelf and logs the time of capture providing a means of visualizing the state of the book shelf at any time.

3.1 Change Notification

Each of the books are labeled with passive RFID tags. Each tag has a unique identifier that corresponds to a database entry containing the name of the book, author and other searchable data. The antenna is placed in an unobstructive place such that the bookshelf is within its detection range. The antenna is attached to the reader using a coaxial cable and is kept out of the way.

The system is notified of changes by the RFID reader which is configured to read all tags in its field and report the ids to a server. The information is relayed using TCP/IP sockets where the server is the system. The information sent is a list of all read tags. The read accuracy is built into the reader and developed by the manufacturer. It reads all tags that are in its field multiple times (30 times per second) and takes care of duplicates and collisions during reading providing an accurate list of tags that are present. The system reader is setup to send this information every 20 seconds to our system which processes the list to determine where a book was added, removed or replaced.

3.2 Course Event Detection

When a new list of books is generated by the reader, it connects to a preset TCP/IP socket and sends the list. The sever accepts the new lists and parses it. It then compares the new list with the current state of the shelf and determines what change event occurred. A new book is inserted if there is no prior record of the tag id. Conversely, a book is removed if the new list does not contain an id. A book is replaced if a known tag id is in the newly generated list. Once the type of event is determined the appropriate visual algorithms are used to update the state of the bookshelf.

3.3 Precise Event Localization

Knowing what event occurred, the current foreground is obtained using background subtraction which incorporates the any projected information as described in Section 2.4. Low thresholds were used (15) for the image subtraction to ensure few false negatives and relying on the planar parallax constraint to deal with the false positives. To reduce computational costs, a region of interest in each camera view was preset and only pixels within were considered as shown in Figure 2.11 Bounding boxes for the foreground are computed using the homography filtering technique (2.4). As noted in Section 2, we assume that the book spines lie on the same world plane and the homography is estimated under this assumption. In our experiments the placement of books did not always follow this constraint. Nevertheless as explained in (2.2) if the depth variation Δd is small as compared to d the error introduced by deviations from the coplanarity assumption are small. In our setup the distance of the spine plane from both cameras was about one meter and we found that for small depth variations from the assumed book spine plane, the detection of spine regions was not adversely affected.

A color histogram of the foreground image within the bounding box is computed. Normalized (r, g) histograms were used as color models for the newly added book spines as they have greater illumination invariance.

Previously, the smallest distance (d^*) between and the histograms of the known books a decision was made about the newly detected region. The χ-squared histogram distance measure for comparing the histograms of the changed region with the book histograms was used. If the distance was less than a threshold (below 0.5 where 1 is a perfect match) for all known books, a "book add" event occurred. As described in Section 2.6, the necessary updates are made to the system to account for the new book.

If d^* was above the threshold a decision between "Book Remove" event or a "Book Replace" event needed to be made which was discussed in Section 2.7. The computed histogram distance for the "Book Remove" event was found to be an order of magnitude higher than for the "Book Replace"event. Based on the detected change the state of the bookshelf was updated along with the time of a change event. With the addition of RFID tags, this event was explicitly known and in the interest of efficiency and accuracy, the vision comparison methodology was no longer used.

One drawback in directly employing the bounding box to represent the book spine in the foreground detection, is that the camera coordinate system may not be exactly aligned with the edges of the shelf. This results in the spine regions obtained by the homography filtering operation to be tapered. As a result the computed bounding box can enclose several "OFF" pixels. If this effect is not accounted for, some regions from the spine of the next book may be included in the model for the current book causing an error that accumulates with each successive addition. One way to deal with this is to estimate the skew of the edges of the shelf and to align the two coordinate systems before computing the bounding box. An alternative approach is to simply keep a record of the "ON" pixels within the bounding box. This mask together with the bounding box information prevents the error from being propagated with the addition of successive books and obviates the need for computing the skew of the shelf edges with respect to the camera coordinate system. The bitmap was used along with (2.5)to update the background model. The mask is also taken into account when computing the color histogram for each change region.

For each of the events that are detected, the foreground representing the state of the shelf has to be generated. For a book add and a book replace event, all masked spines of books are drawn on the original background of the empty shelf. For a book remove event, the detected foreground is replaced with background pixels of the empty shelf. For a book add event, if all the tags are not pre-scanned, the user has to enter the user name and title of the book. For book remove or replace events, since the books were previously in the system, the book is placed in either the not present or the present system state list respectively.

To access a particular book the user enters its title or author. By performing a lookup of the state lists, the bounding box information along with the bitmap related to that book(s) is obtained. The coordinates of the ON pixels are warped using the homography between the projector and the camera and the desired book is highlighted as shown in the Figure 3.1(a). Additional information about the book can be projected on the upper edge of the shelf as well.

Since our system maintains a record of camera images representing the states of the bookshelf when changes took place, we can rewind and display the state of the bookshelf at any time in the past by displaying the camera image of the time on or the last record before the stated user time. The images could be warped and projected to get an easy access to the desired state of the bookshelf as shown in Figure 3.1(f).

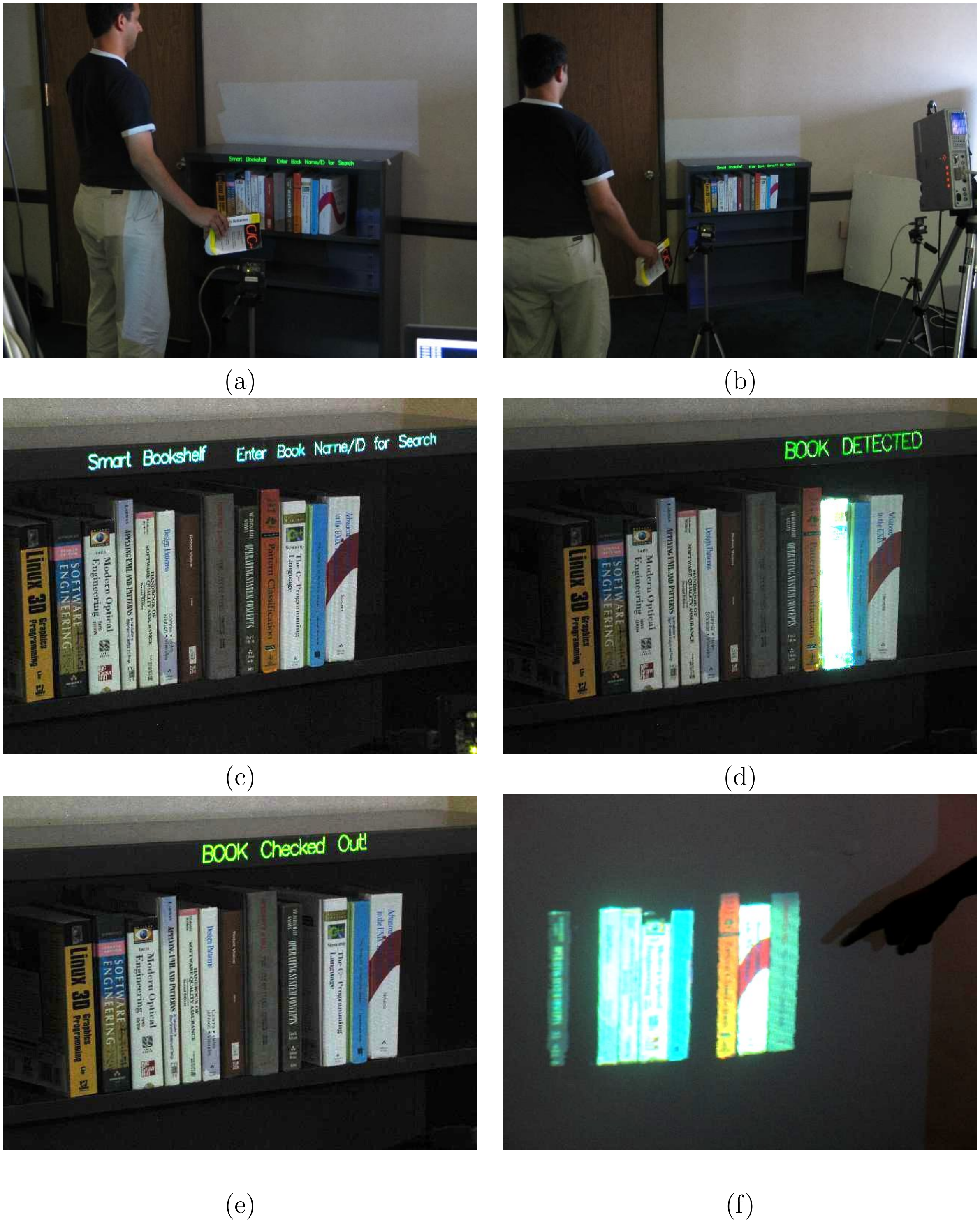


Figure 3.1: Images of the smart bookshelf (a) and (b) the setup (c) (d) and (e) show different states of the bookshelf in response to a query by the user (f) shows the projection of the requested past state of the bookshelf.

Chapter 4

Conclusions and Future Work

This work presents an initial step towards environments that are intelligently augmented by a camera-projector systems. The ”smart bookshelf” uses two cameras and a projector, to provide coverage of the bookshelf to be monitored. The system builds a database that contains information related to book position, time of addition and the texture information of the book spine. This multimodal system incooperates an RFID reader along with the vision system to continuously monitor the bookshelf to detect addition, removal and reinsertion of books. Accurately knowing the events that occurred, the system maintains the inferred state of the bookshelf. The two pronged event detection approach utilizing RFID information to trigger the visual system seems to be an improvement over pure vision approaches we initially explored and reduces false positives.

The foreground detection algorithm which takes into account the projected information along with a planar parallax constraint is used to extract the spine of the book. The idea of geometric transfer has been used for other tasks such as obstacle avoidance and face recognition and was the key component of our approach. In environments where illumination is changing dramatically (via a projector) geometric verification is important. Based on the detected change, the state is updated. Knowing the homography induced by the spline plane between the projector and the cameras, the smart bookshelf supports user queries for the presence of a book, by automatic highlighting the book on the shelf or indicating its absence if has been removed. Furthermore the system maintains a record of the state of the bookshelf which can be replayed as desired.

At present our approach assumes that a single change event occurs over a fixed time interval. The primary reason for the inclusion of RFIDs was to relax this constraint as it has the ability to read several tags at once. However, we quickly realized that the RFID component could be used the enhance the system. It was used to make event detection more robust by using as a triggering mechanism which would inform the visual system of what event had occurred and to update the state of the system accordingly. It also increased the usability of system as it allowed us to associate book information with a tag id. The latter would not require the user to enter book information when books were added to the system as the books could be pre-scanned and the information could be stored separately. Additionally, if a unscanned tag was introduced this information could be entered once and stored permanently

The RFID addition has given us the ability to address the issue of multiple additions and removals which was previously not possible using only vision. Removals would be indicated by the reader as tags not being present in the RF field and no visual cues would be needed to update the state of the system. Multiple additions however, fall into two cases. The first workable case would be addition of multiple books that have known appearance models (books that have previously been in the shelf). The known appearance could be used as a template which could be used to determine the new location of the book in the changed foreground region. The second case is when new books are added to the shelf. With no appearance models to distinguish between the newly added books, this case cannot be solved. However, knowing that this event has occurred would greatly increase the usability of the system as we now have the ability to inform the user that this has happened and cannot be handled at the current time. Such a situation would require user intervention to be resolved. When books were added and removed, user had to take care not to drop the other books. Or, overtime if the books accidentally fell, this would be visually interpreted as a multiple addition or removal. However, the RFID system would inform the system that a single addition, a single removal or no change in state occurred respectively. The system now has the ability of take care of book fall events.

4.1 Extensions

More research was done to increase the reliability of the homography filtering technique which was used for foreground detection. It was observed that the technique does not effectively capture pixels of the side of the book that lie on the base plane, which is orthogonal to both the spine plane and the planes that contain the sides of the book. The ongoing work that is being conducted so that we can better classify pixels that lie on the spine plane and not on the other planes. Knowing the other planes and utilizing the homography filtering technique, we can classify pixels that correspond to the spine plane only. Through cross validation, we can effectively rule out foreground pixels of the other planes, robustly extracting spine pixels without the need for unreliable color matching. Determining homographies of the the other planes can be done using

$$H_2 = H_1 + e' l_{12} \tag{4.1}$$

[6] where H_1 is the homography of the spine plane; e' is the epipole of the left camera and l_{12} is the intersection line of the H_1 and either of the two orthogonal side or base planes. Theses lines could be determined from the images of the book shelf itself which has these edges clearly visible due to it structure. Since placement books in the shelf could span from left to right and bottom to the top, we would have to find intersecting lines with the spines planes at those cuts. This would allow us to determine min and max planes for the side and base planes respectively. Having those planes, we could find the 'tween planes by interpolation

$$H_{sidemin} = H_{spine} + e' l_{left} \tag{4.2}$$

$$H_{sidemax} = H_{spine} + e' l_{right} \tag{4.3}$$

$$H = (1 - \alpha) * H_{sidemin} + (\alpha) * H_{sidemax} \tag{4.4}$$

where l_{left} is the line intersecting the spine plane and the left of the bookshelf;l_{right} is the line intersecting the spine plane and the right of the bookshelf and α is the center image point of the detected bounding box in the side plane. Similarly we could estimate the homography corresponding the base plane and accurately determine spine plane pixels. These pixels would be "on" pixels in the foreground image that use the spine plane homography and not the other planes.

Having a notion of the planes that foreground pixels lie in, it would be interesting if we could infer positions and orientation of books in the shelf. Also there is a lot of interesting work from the field of structured light that can be used for this application. The projector could be used not only as communicative medium, but possibly enhance the systems performance to increase reliability. However, that is work for a later time.

4.2 Future Directions and Impact

Computer vision tightly coupled with camera-projector systems enable a variety of exciting augmented reality applications. Work such as [1] and [14] allow human computer interaction to become more tangible, more natural. Computers have become fast enough so as to interact with the humans in terms that we are more used to. The interaction will seem more intuitive and it will become more ambiguous. We will interact with virtual things in a non-virtual way. The two systems described above are geared for single users in a limited space. With their continually dropping price point, it has become feasible to deploy cameras and projectors everywhere allowing these systems to span wider areas and larger spaces. There will be cameras observing and pixels everywhere we see. Applications such as these can be used for assisted living for example. They can be deployed in smart households and help with disabled or old people who have mobility issues. Help in the form of projected information can come to them anywhere a projector and camera is present. Conversely, in a manner similar to how head mounted displays brought us into the computer world, cameras and projectors will bring computers into ours. Wide area systems such as the ambient virtual assistant (AVA) which is part of the EPSCoR grant could one-day be the norm. These envision environments where the physical world acts as the interface to our computation infrastructure and every surface is a potential display for our information. Interactive applications will become interactive spaces, houses and buildings. As the technology scales, so will the size, giving way to interactive spaces everywhere.

This work has served its purpose in that it has been a testbed were we one can determine the issues of building an intelligent environment by actually creating one. Although the system is still far away from being commercially viable, the lessons learnt have been invaluable. Being humans, we take for granted all that we can see and do because it is second nature to us. I believe that computer vision and applications such as this are still in their infancy, but on their way to becoming as reliable and important to our experience as the human visual system.

Bibliography

[1] A.Kale, K. Kwan, and C. Jaynes, *Epipolar constrained user bushbutton selection in projected interfaces*, Proceedings of the 1st CVPR workshop on real time vision for human computer interaction, (2004).

[2] P. Batavia and S. Singh, *Obstacle detection using color segmentation and color stereo homography*, Proceedings of ICRA, (2001).

[3] S. Borkowski, O. Riff, and J. L. Crowley, *Projecting rectified images in an augmented environment*, Proceedings of PROCAMS, (2003).

[4] R. R. et al, *The office of the future : A unified approach to image-based modeling and spatially immersive displays*, Proc. of the ACM SIGGRAPH 1998, (1998).

[5] R. R. et al., *ilamps: geometrically aware and self-configuring projectors*, ACM Transactions on Graphics, (2003).

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

[7] C. Jaynes, W. B. Seales, K. Calvert, Z. Fei, and J. Griffioen, *The metaverse- a networked collection of inexpensive, self-configuring immersive environments*, 7th International Workshop on Immersive Projection Technology, 9th Eurographics Workshop on Virtual Enviroments, (2003).

[8] C. Jaynes, S. Webb, M. Steele, M. Brown, and B. Seales, *Dynamic shadow removal from front projection displays*, Proc. of IEEE Visualization, (2001).

[9] A. Majumder and R. Stevens, *Color non-uniformity in projection based displays*, IEEE Transactions on Visualization and Computer Graphics, (2003).

[10] G. E. Moore, *Cramming more components onto integrated circuits*, Electronics Magazine, (1965).

[11] S. Nayar, H. Peri, M. D. Grossberg, and P. N. Belhumuer, *A projection system with radiometric compensation for screen imperfections*, Proceedings of PROCAMS, (2003).

[12] C. Pinhanez, *Creating ubiquitous interactive games using everywhere displays projectors*, Proc. of the International Workshop on Entertainment Computing, (2002).

[13] W. B. Seales and J. Caban, *Visualization trends: Applications in laparoscopy*, Seminars in Laparoscopic Surgery, 10 (2003).

[14] C. J. Shilpi Gupta, *Active pursuit tracking in a projector-camera system with application to augmented reality*, in Procams, June 2005.

[15] S. Webb and C. Jaynes, *The dome: A portable multi-projector visualization system for digital artifacts*, in IEEE Workshop on Emerging Display Technologies (w/ VR 2005), March 12-16 2005.

[16] P. Wellner, W. Mackay, and R. Gold, *Computer-augmented environments: back to the real world*, Communications of the ACM, (1993).

Vita

Background

- Date of Birth: October 17th 1979
- Place of Birth: Mumbai, India

Education

- Master of Science in Computer Science, University of Kentucky, 2006
  Anticipate completion and defense of thesis in April.
- Bachelor of Science in Computer Science, University of Kentucky, 2002

Professional Experience

- October 2005 - Present. Research Engineer, Center for Visualization and Virtual Enviroments.
- January 2005 - October 2005. Research Assistant / Assistant System Administrator. Center for Visualization and Virtual Environments.
- September 2003 - December 2004. Web Developer, Programmer/Tech Support. University of Kentucky, Lexington, KY. (College of Design).

Publications

- D. Crasto, A. Kale, C. Jaynes, The Smart Bookshelf: A study of camera projector scene augmentation of an everyday environment, IEEEs Workshop of applications of Computer Vision, Breckenridge, CO, January 5-7 2005.