Sensors and Signal Processing For High Accuracy Passenger Counting

Final Report

Scott E. Budge
scott.budge@ece.usu.edu
435-797-3433

John Sallay
j.sallay@aggiemail.usu.edu
435-797-0464

Department of Electrical & Computer Engineering
Utah State University
4120 Old Main Hill
Logan, Utah 84322-4120

5 March, 2009

1 Introduction

It is imperative for a transit system to track statistics about their ridership in order to plan bus routes. There exists a wide variety of methods for obtaining these statistics that range from relying on the driver to count people to utilizing cameras and sensors. Utah State University (USU) has undertaken the task of creating a high accuracy people counter using a texel camera. The project has two main objectives. The first and most important is to develop a system that accurately counts the number of people entering and exiting a bus. The second is to associate each exiting passenger with a passenger that previously entered. This information will better allow a transit system to track usage of buses and stops.

The project has been divided into two phases. This report covers the work done in phase I. Phase II would be carried out under additional funding. The primary goal of phase I is to develop a proof of concept system for the people counter with the software algorithms needed to perform accurately. Phase II will involve the optimization of the algorithms used and the design and manufacturing of a prototype system that can be installed a bus. The optimization will be performed at USU, while VPI Engineering (a third-party interested in commercial system production) will work to produce a prototype system.

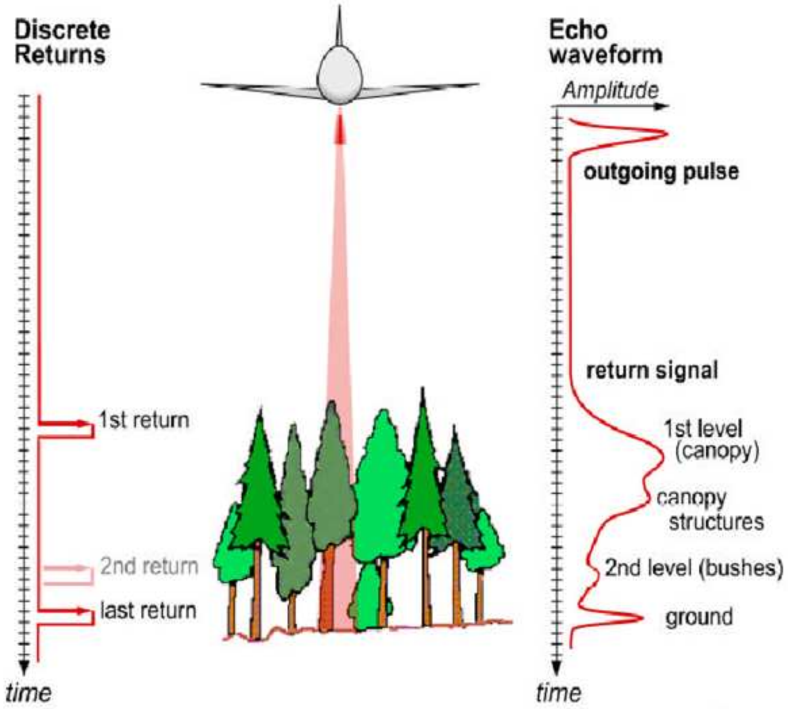


Figure 1: Lidar Example.

This report is organized as follows: Section 2 gives an overview of flash lidar and texel technology with comparisons to other current technologies, Section 3 details the algorithms developed to implement the people counter, Section 4 describes the preliminary results found, Section 5 presents the good-faith cost estimate produced by VPI Engineering, and Section 6 contains our conclusions.

2 Flash Lidar and Texel Technology

A digital camera receives incoming light and describes a scene with color information. It is a simple task for a human being to separate objects in a digital image, but it can be quite a difficult task for a computer to undertake, especially when the objects might be overlapping or of the same color.

A Light Detection and Ranging (lidar) sensor measures the time it takes for a pulse of light to travel from the lidar sensor, reflect from an object, and return to the sensor. Using this information it determines the distance of objects from the lidar as is illustrated in Figure 1. As shown, if the light reflects from multiple surfaces, it is possible to find the distance from the closest surface (the tree canopy), surfaces in between (bushes), or other surfaces along the path of the light pulse. Usually the first return is desired. Lidar is commonly used in mapping ground geography, target identification, and obstacle avoidance. If the lidar is capable of measuring the distance from several locations in the scene with a single pulse, the lidar is called

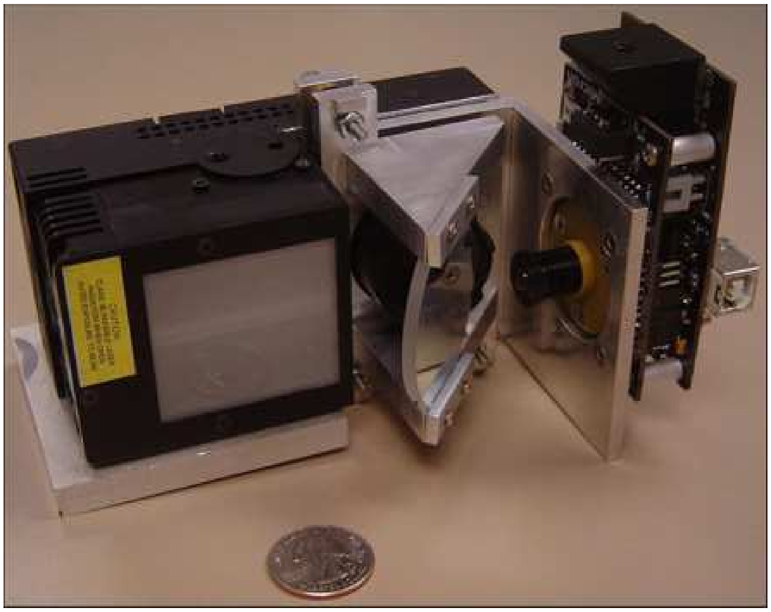


Figure 2: Prototype Texel Camera.

a flash lidar and the result is a "depth image" of the scene.

Oftentimes it is quite difficult to separate people from each other or background objects in a color image, but the depth information produced from the flash lidar lends itself naturally to the task. As a person enters a scene there is a significant change in the depth information that the lidar receives from the background and the person with each snapshot. It then becomes a relatively simple task to separate the person from the background.

A texel camera is the combination of a flash lidar with a digital camera, producing both depth and color information from a single snapshot. The two cameras are mounted together, 90° apart. Incoming light is intercepted by a cold mirror, which allows the infrared light pulse transmitted from the lidar to pass through to the lidar sensor on its return from the scene, and reflects visible light from the scene into the digital camera. In this way both cameras receive light from the same field of view. The texel camera fuses the depth information spatially with the color information, producing a 3-D representation of a scene. The texel camera used for this project is shown is Figure 2.

Texel technology provides several advantages over other possible technologies that could be used in people counters. First, background separation is more effective as described above. Second, it can perform better than light intensity-based systems when the lighting on a bus varies throughout the day. Shadows, weather, and bright lights are common situations that a people counter would encounter. A light intensity-based imager might not perform well through these changing circumstances, as compared to a lidar system. Specifically, the lidar transmitter is made up of an array of monochromatic modulated LED sources and the receiver is primarily sensitive to the LED wavelength. Since the light is in a very narrow band of wavelengths, the lidar will not be affected much by changes in overall lighting.

There is more information available in a texel image than would be available in a digital image. This is especially important for the task of recognizing when a specific passenger enters and leaves the bus. The color information in conjunction with the depth information allows us to better match an exiting passenger to a database of passengers currently on the bus. The ability to identify the characteristics of a person will also allow us to distinguish better the difference between a person and a large object the person is carrying, such as skis or a duffel bag. In this manner, the task of recognizing people re-enforces the task of counting. This will allow a transit system to know where individual passengers get on and off a bus which will be very useful in route planning.

3 Processing Algorithms

In order to reduce the complexity of a texel image, the camera is mounted above the bus door looking down at the floor of the bus. The texel images are therefore acquired from above the passengers entering and exiting the bus. This allows a consistent view of the passengers, makes mounting the camera in existing buses easier, and reduces potential camera damage caused by bumping the camera.

The processing problem can be broken down into three tasks:

1. Track people as they move through a frame.

2. Count people as they enter and exit a bus.

3. Match a person leaving with the same person that previously entered.

Before any of these tasks can be completed, some preliminary steps must be taken. A depth and color image are captured simultaneously. Both images will have some distortion due to imperfections in the lenses of the cameras. The distortion is determined through a calibration procedure for the lidar camera and a correction is applied in hardware or software. The color image is then mapped to the corrected depth image. The combination of the two is the texel image used.

3.1 Track people as they move through a frame.

The first step in tracking people is noise and background removal. These operations are performed using just the depth image because it is far easier to distinguish the background in a depth image than in a color image. The image is thresholded to remove any bad pixels and background. For example, it is common to have a few noisy pixels with a value larger than the distance to the floor of the bus. Any pixels that are near the distance to the floor of the bus are considered either background or error measurements. These pixels are removed from the image. The image is smoothed using a median filter to fill any small holes created by the thresholding. All of the objects in the image and their sizes are then found. An object is defined as a group of connected pixels. Any object that is too small to be a person is removed from the image. This creates a much cleaner image that only contains the objects that we are interested in tracking.

As people move throughout the image it is necessary to track them. This is a trivial task with single persons in the field of view, but it can become quite complicated when there are multiple people in a frame. It also becomes difficult as people enter and exit. One person could exit the frame at the same time that someone else enters.

After noise and background removal, the current frame is subtracted from the previous frame to form a difference image. If there is very little difference between the two frames then very little motion occurred between the frames. A large difference either means that someone moved significantly or a new person has entered the frame. The location of the movement in the difference image and the direction of travel of the person in the frame distinguish between the two. A new person will enter near the edges of the frame while someone in the image from one frame to the next can't change their direction or speed significantly in one frame time.

Sometimes it may be difficult to distinguish between two people. When two people run into each other it can be very difficult to determine where one person ends and the other begins. A motion prediction algorithm is used in these situations. We have chosen to use an algorithm very similar to that used in MPEG video encoding for exploiting motion between frames. We again use the previous and current frames. There is no information about where a new person was before entering the frame; thus, any new people are removed and the motion prediction is only performed on those who were in the previous frame.

First, the previous frame is divided into 8x8 pixel blocks and a search window is created in the current frame corresponding to each block. We assume that a block can't move a significant amount between frames. The search window size is determined by the amount of motion that would be reasonable to occur, and its location is determined by predicting where the block would move given the velocity of the block in the previous frame, as illustrated in Figure 3.

Each block is shifted to every offset (s, t) in the corresponding search window and the match error is found according to

$$e(s,t) = \sum_{i=1}^{N} \sum_{j=1}^{N} |cp(i+s, j+t) - pp(i,j)| \tag{1}$$

for both the depth and the color pixels, where $pp(i, j)$ is a pixel from the previous image, and $cp(i + s, j + t)$

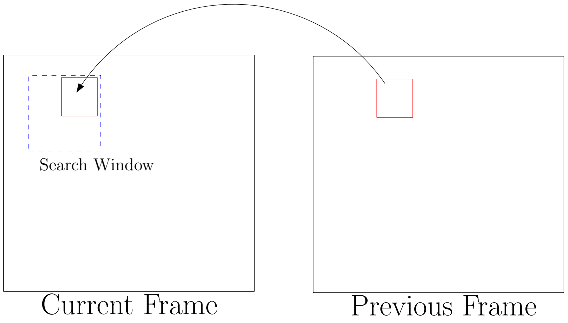
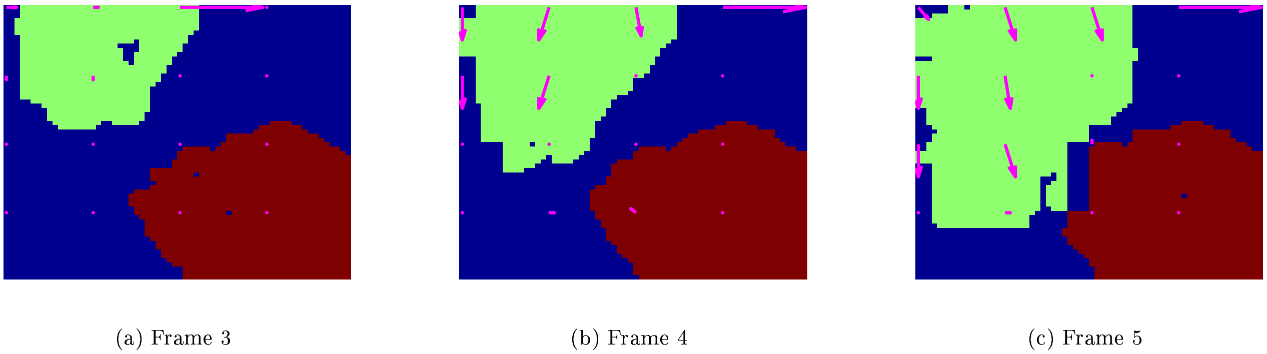


Figure 3: Motion prediction search from one frame to the next. The



(a) Frame 3

(b) Frame 4

(c) Frame 5

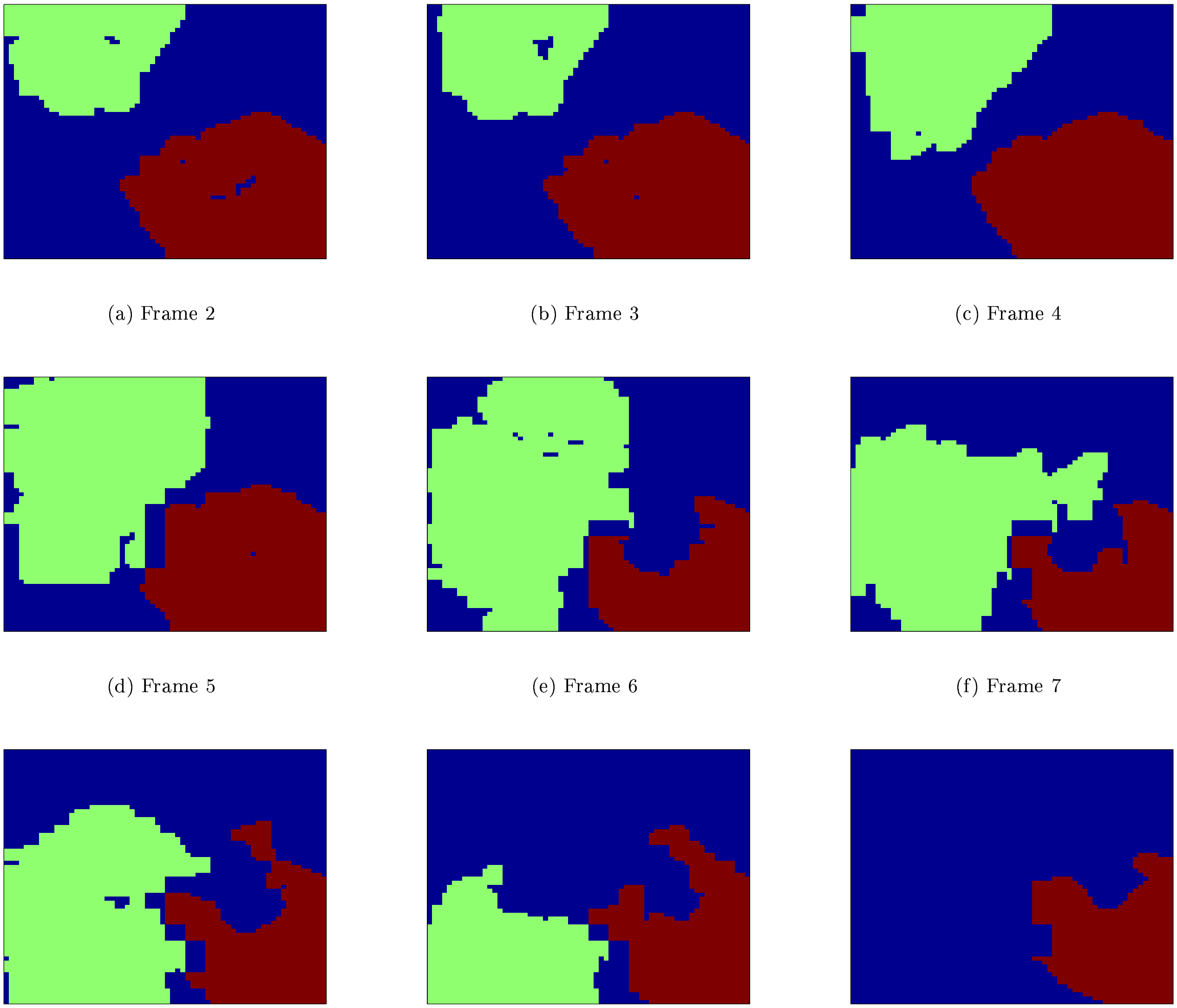
Figure 4: Motion vectors for two successive frames.

is a pixel from within the search window in the current image. The value of (s, t) that minimizes the error is used for the motion vector for that block. The velocity of a person is then calculated by taking the average of all the motion vectors of the blocks associated with that person.

There may be some pixels that aren't associated with the previous frame. As a person enters only half of their body may be in one frame and their whole body in the next. The back half of the person isn't in the first frame and thus won't have any motion vectors associated with it. The pixels with no associated motion vectors are assigned to a person based upon the values of the neighboring pixels. In this manner every pixel in the current image is mapped to a person or background from the previous frame.

This algorithm compares the current frame and the previous frame so that a person can be tracked from frame to frame. Figure 4 shows the motion vectors from three successive frames where the green blob is a person moving across the field of view past a stationary person represented by the brown blob.

The segmentation is good enough to track the motion of persons in the images. Figure 5 illustrates the performance of the algorithm when one person passes another in the frame. Note that the persons represented by the blobs touch each other in passing, but remain distinct.



(a) Frame 2

(b) Frame 3

(c) Frame 4

(d) Frame 5

(e) Frame 6

(f) Frame 7

(g) Frame 8

(h) Frame 9

(i) Frame 10

Figure 5: Motion Prediction Sequence

3.2 Count people as they enter and exit a bus.

When a person enters and exits a frame, the side they enter from is recorded. If they start in the top half of a frame they are considered entering the bus. If they start in the bottom half they are exiting the bus. When a person exits the frame, the two sides are compared. A person could get half way onto the bus and then walk back off. They should not be counted in this situation. If the two sides are different, the person will be counted. Depending on the direction of travel, the count of people on the bus will be incremented or decremented.

3.3 Match a person leaving with the same person that previously entered.

There exist many ways to perform this association. One method that is commonly used is known as matched filtering. A template of a person's image is created when they enter and stored in a bank of templates for everyone on the bus. When a person exits they are compared to every template in the bank. The best match is chosen and removed. The disadvantage of using a matched filter is that the template does not perform well under rotational changes. It would not be uncommon for a person to enter and exit at a different angle. It would also require a lot of storage and computations to match a person's image template.

The solution chosen is similar but with some important distinctions. The following features are collected for every person who enters the bus:

1. Height
2. Hair Color
3. Hair texture
4. Shoulder height
5. Shoulder color
6. Shoulder texture

If a person rotates, these features should remain approximately invariant. Features are collected for every frame that a person is in the field of view of the camera. As soon as the person exits the field of view of the camera, the features are averaged together to reduce noise in the measurements. If the person is entering the bus, the features are used to construct a "feature vector" and the vector is added to a feature bank. If the person is exiting the bus the feature vector is compared to all of the other vectors in the feature bank and the distance to each stored vector is computed. By using feature vectors, identifying information is still collected, but it requires less memory and fewer computations to match a person.

Data	#enter	#counted	#exit	#counted	percent
Quick	9	9	12	11	95%
Bright	0	0	5	5	100%
Collide	5	5	4	3	89%
Together	3	0	5	0	0%
Other	52	51	43	43	99%
Total	69	65	69	63	92%
Adjusted	66	65	65	63	98%

Table 1: Counting performance

The distance is found by using the Euclidean distance between the features, given by:

$$d_{i,j} = \sqrt{\sum_{k=1}^{n} \left[F_i(k) - F_j(k)\right]^2}, \qquad (2)$$

where F_i is the feature vector for the exiting person, F_j is the feature vector corresponding to the j^{th} person in the feature bank, n is the total number of features, and $F_j(k)$ is the value of the k^{th} feature for the j^{th} person. The vector F_m that produces the smallest distance is considered the best match and is removed from the feature bank.

At this point many useful statistics can be gathered. The simplest and most important is the count of how many people used the bus. The number of people who entered and exited at each stop can also be counted. Since we also know *who* entered or exited the bus, we can determine the length of time each person was on the bus.

4 Experimental Results

Test data was collected to analyze the accuracy of our algorithm. Our results are summarized in Table 1. Data was collected on a bus during our status report in October and in our lab at USU over a period of several months. Several scenarios were tested. These include:

1. Persons entering quickly.

2. Persons entering in bright sunlight.

3. People running into each other.

4. People entering together (i.e. arm around each other)

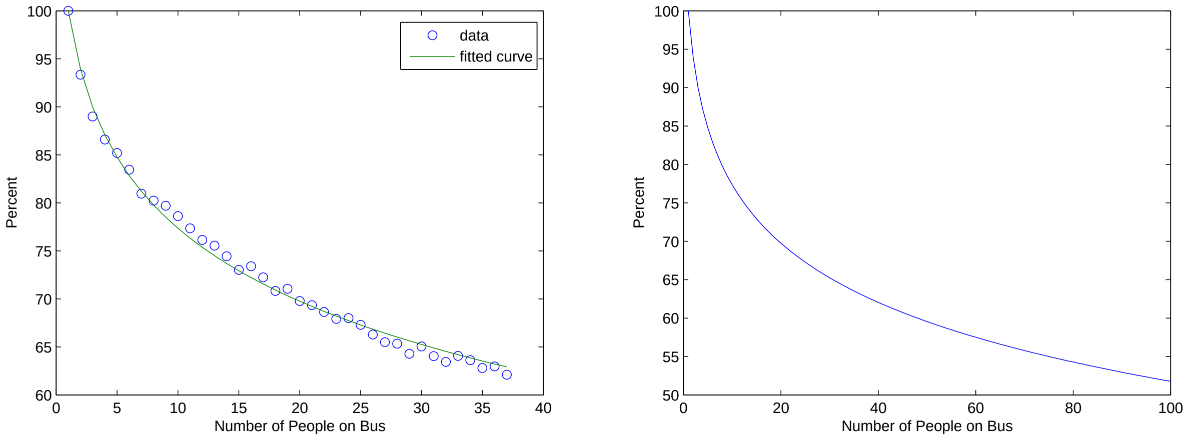
The system works very well when people enter quickly. We were able to achieve 95% accuracy with 21 entrances and exits. Direct sunlight did not have any effect on our data and we achieved a counting accuracy of 100% correct. Collisions did not cause a significant counting problem. The motion prediction algorithm described in Section 3 was able to separate the different people very well and we correctly counted in 89% of the cases tested. The greatest difficulty occurs when people enter together. Our system is not currently able to distinguish between them well and we were not able to count correctly in any of these situations. Most of the missed counts are due to this issue. The rest of the data collected did not fit into any of these situations and we achieved a counting accuracy of 99%. This produces an overall counting accuracy of 92% correct.

We feel that the difficulty with counting people who enter together is more of a hardware limitation than an algorithm limitation. The resolution of the lidar camera we are currently using is about half of that of newer cameras, resulting in a reduced ability to distinguish the depth changes between people. In addition, the depth and digital images also are not taken at exactly the same time due to software triggering. When there is a significant amount of motion between frames there can be large errors in registration between the digital and the lidar images. It is difficult to distinguish between people when the images are not aligned in time. Both of these problems could be reduced in the next hardware prototype with a higher resolution depth camera and hardware triggering for frame capture.

The adjusted count given in Table 1 removes the "Together" data resulting in a correct count in 98% of the situations tested. We feel that this is a more accurate representation of the system's capabilities with improved hardware.

Matching accuracy is difficult to measure. It is far easier to perform the correct association when there are 2 people on the bus than 80. In addition, the order in which people leave can greatly affect the results. For example, a matching error causes an incorrect person to be removed from the feature bank. When that person leaves another error will occur because their set of features is no longer in the feature bank. We therefore attempted to understand the matching performance of the system using Monte Carlo analysis.

The matching performance of the system was estimated using a database of 37 individuals. Feature vectors were gathered for each of those 37 people as they entered the "bus." A subset was selected at random to simulate a set of people on the bus. A single person in the set was randomly selected as the person exiting and a feature vector for this person (the "exiting vector") was created from data obtained during an exit. Note that this is a different vector than was created for the database of entering persons. Next, the matching algorithm was performed which compared the exiting vector to the subset. The experiment was repeated 10,000 times and the percentage of correct matches was computed. Figure 6a shows the accuracy for different numbers of people on the bus. The accuracy decreases with the number of people on the bus. With 37 people on the bus we were able to achieve an accuracy of 62%.



(a)

(b)

Figure 6: Results for person matching. (a) Matching accuracy using experimental database. (b) Extrapolated results for larger numbers of riders on the bus.

The data collected matches a logarithmic function,

$$f(n) = Alog(n - \alpha) + b, \tag{3}$$

where $A = -11.3$, $\alpha = 0.4$, and $b = 103.8$. This function was used to extrapolate the data out to 100 people as shown in Figure 6b.

The function decreases very slowly over the latter half of the curve; thus, there should not be a significant difference in matching accuracy as the number of people on the bus increases. Even with 100 people on a bus the model predicts an accuracy of over 50%.

There is still work to do to improve our results. The matching accuracy suffers from the same problems described for counting. In addition, we have not investigated more sophisticated algorithms which are available for matching patterns. We expect to make improvements in our algorithm and in our hardware that will enable us to achieve 80% accuracy with 70 people on a bus.

In addition, we have not investigated methods to correct for errors in matching. Techniques based on soft decision-making or the closeness of feature vectors in the decision space can be developed in the optimization phase of the project.

5 Estimate of Cost

Before we move on to Phase II of the project, UTA has requested that we provide a cost estimate for production, installation, and maintenance of people-counting systems on buses. USU was able to team with a local Utah company, VPI Engineering, to provide a good-faith estimate of these costs. As a first step in this estimate, VPI has provided two conceptual designs, one for a prototype system, and the second for a production system. The prototype system is for developing the people-counter concept into a ruggedized hardware design for testing on a bus, while the production system is intended to be the lower-cost design for deployment throughout the fleet of buses.

An example of the prototype conceptual design is given in Figure 7. It seeks to lower development time by using commercial-off-the-shelf (COTS) components which can be assembled together with a minimum of custom design. The estimate of costs for this design is based on the construction of five prototype units for testing on UTA buses.

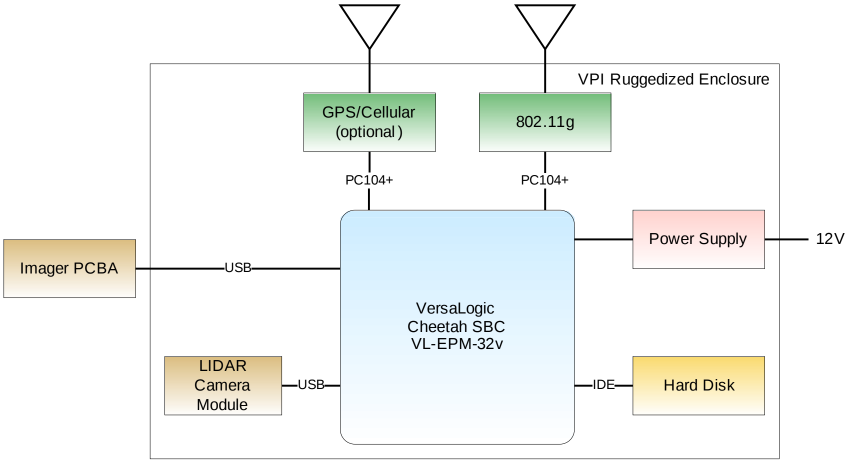


Figure 7: System prototype design.

Once the prototype units have been installed and tested, a production system will be developed, as given in Figure 8. The conceptual design includes a higher level of integration, which should result in better performance, smaller size and weight, and lower cost per unit.

Rough-order-of-magnitude (ROM) estimates for developing this technology into a deployed system are given in Table 2. VPI has divided the development into two phases. Phase I entails the manufacture of the prototype units and Phase II is for the development of the production unit. The prototype development requires a lower non-recurring engineering (NRE) cost, with higher per-unit cost. Once this phase has been completed, the NRE cost of the final design is higher, reflecting the cost of custom integration. The final per-unit cost is much lower, estimated to be slightly more than $4,000, based on a total of 100 units. The cost per unit including NRE works out to be $5,700 per camera. Each of these estimates is based on an aggressive development schedule of about six months from start to completion of the production model. Adjustment of the development schedule may result in lower costs.

VPI has also provided a preliminary cost estimate for maintenance for the systems. This estimate includes

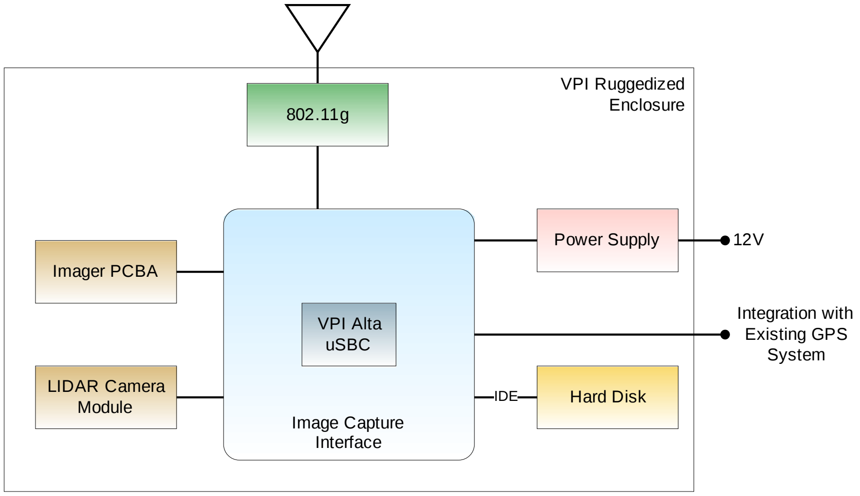


Figure 8: System production design.

Description	**Labor**	**Expenses**	**Total**
Phase I Non-recurring Engineering Costs			$76,000
Phase II Non-recurring Engineering Costs	$160,000	$12,000	$172,000
	Qty	**Unit Cost**	
Prototype Unit	5	$14,000	$72,500
Production Unit	100	$4,050	$405, 000

Table 2: Estimated Costs for System Deployment

two options:

- Option 1: 90 day warranty, full replacement during warranty period, included in unit cost. No annual maintenance contract, any work on systems provided on time and materials basis. Hourly rate of $70/hour plus material expenses.
- Option 2: Maintenance contract, quarterly or annual maintenance fee. Quarterly maintenance inspections, full warranty replacement, software upgrades included. Annual maintenance fee of 30% to 35% of total order cost. Maintenance agreement to last 4 years. At end of 4 years, review options to upgrade with replacement technology.

6 Conclusions

We have demonstrated that the use of a texel camera-based people counter is an effective method for accurately counting the number of persons entering and exiting a bus. A database of 37 individuals entering and exiting the field of view, acquired under differing conditions, was tested for counting accuracy and our system was able to achieve over 98% correct counts. We feel that this represents a high degree of accuracy and can significantly improve the ability of UTA to know the load on buses and the instantaneous count of persons on an individual bus in the event of an emergency.

In addition, the system was tested for its accuracy in determining which individuals enter and exit a bus at a particular stop. Since this is difficult to quantify, we performed a Monte Carlo analysis of the database of 37 individuals. Our resulting estimates indicate that the system is better than 60% accurate in correctly matching passengers leaving the bus carrying 37 people. With smaller numbers on the bus the performance improves. The data seem to support an extrapolation that over 50% will be correctly identified in a database of up to 100 passengers.

Our experience with the system has also identified several approaches to improve the performance of the system. Current technology in lidar (depth) sensors is more than twice the resolution of the camera used in this study. In addition, better optics and synchronization of the digital camera to the depth image should result in more discriminating passenger matching. A second direction for system optimization is to investigate more sophisticated methods for performing the classification itself. Finally, methods for handling incorrect matches can be developed.

Finally, our good-faith estimate of the costs of the system lead us to predict that we should be able to install the system into UTA buses for about $5,700 per camera.