Multipedestrian Tracking

by David R. P. Gibson, Bo Ling, Michael Zeifman, Shaoqiang Dong, and Uma Venkataraman

Public Road Magazine - March/April 2006 issue

A new detection system using computerized stereovision promises greater pedestrian safety in the years ahead.

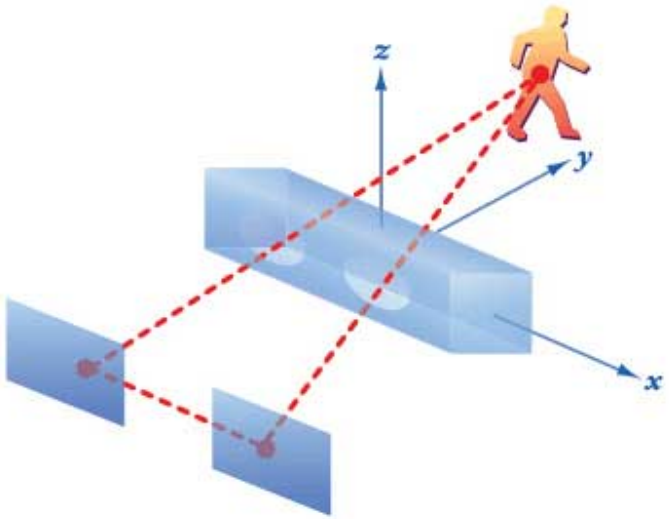
(Above) New research at FHWA to track pedestrians like these crossing a street in a crosswalk offers the promise of reducing fatalities and injuries. *Photo: AAA Foundation for Traffic Safety*

Every year, nearly 5,000 pedestrians are killed in traffic incidents in the United States. To improve safety in the roadway environment, researchers at the Federal Highway Administration (FHWA) are applying the latest technologies to detect and track pedestrians in crosswalks. Accurate computer tracking of pedestrians offers the possibility of developing in-vehicle instruments for helping motorists detect and avoid individuals in crosswalks. It also could lead to provision by the traffic control system of safe "walk" and "don't walk" clearance periods for pedestrians.

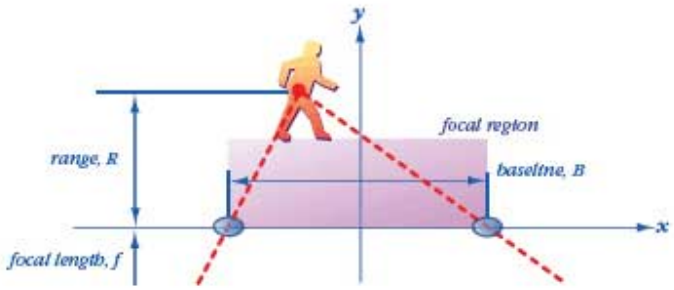
A five-member team of FHWA researchers and contractors deployed a set of two video cameras aimed at an intersection from various angles, using the cameras and a computer to record the movements of pedestrians. An important component of the research was to distinguish multiple people from each other and the intersection's background environment. The system also estimated the pedestrians' walking speed and their location as they traversed the intersection.

"We need to find ways to know that pedestrians are at intersections and need to be served," says Tom Dodds, P.E., pedestrian and bicycle engineer, South Carolina Department of Transportation (SCDOT). "I have noted here in South Carolina that pedestrians are not always particularly amenable to locating and pushing the pedestrian button. It would be more critical to locate

pedestrians during nighttime as opposed to daytime because drivers have a more difficult time seeing and perceiving pedestrians [at night]. Nighttime conditions, along with various foul weather scenarios, are critical for pedestrian safety."



The diagram shows the basic stereo camera formation in three dimensions, *x*, *y*, and *z*. The rectangular box represents a stereo camera with two lenses. The other boxes represent the focal planes within the camera, and the plane outlined by the dashed triangle is called the epipolar plane. *Source: Migma Systems, Inc.*



This figure shows a 2-D view of a stereo camera formation. The line between the right and left focal centers is called the baseline, *B*. In general, *B* is quite small. For instance, the baseline is just 12 centimeters (4.7 inches) in the stereovision system set up for FHWA's research. The distance between the image plane and focal center is called focal length, *f*. The area between right and left focal rays is the focal region. The factor of most interest is the distance between the object and

stereo camera baseline, often called range, *R. Source: Migma Systems, Inc.*

In addition to offering the potential for reducing injuries and fatalities at intersections, the FHWA line of inquiry may soon be reflected in new in-vehicle instrumentation or other technologies that could reduce vehicle-pedestrian crashes.

Need for Passive Detection

William C. Kloos, P.E., signals and street lighting division manager for the city of Portland, OR, lays out the case for passive detection of pedestrians: "Accurate and reliable detection of pedestrians and bicyclists is a key to providing safer and more efficient traffic signal operations," he says, citing various reasons:

"At several locations in Portland, pedestrians will push the pedestrian button, but if a gap appears in traffic, will cross against the signal. Shortly after that, the pedestrian phase comes up, but no pedestrian is there. Motorists wonder why they are stopping. If we had reliable pedestrian detection, we could cancel the call [because] the pedestrian phase is served."

Kloos continues: "At many locations, pedestrians don't realize that they have to push the button to get the pedestrian phase. With reliable passive detection of pedestrians, the button would not be needed. We currently use passive detection to extend the pedestrian clearance interval for pedestrians. This method allows us to program the minimum pedestrian clearance time but extend that time as needed for slower pedestrians.

"Passive detection of pedestrians can improve our service to pedestrians with special needs. At many existing intersections, the push buttons may be located some distance from the curb. This location issue is difficult for people in wheelchairs and for people who have low vision. The bicycle mode shift is continuing to increase. With improved bicycle detection we could improve the timing of signals to meet the unique performance needs of cyclists."

Regarding the possible benefits derived from passive detection for those with low vision, the United States Access Board recently released its latest version of the *Draft Guidelines for Accessible Public Rights-of-Way,* which will require the installation of accessible pedestrian signals at all new signalized pedestrian crossings.

Challenges In Computerized Pedestrian Detection

Vision-based pedestrian detection in an outdoor environment is a challenge. Although it may be simple for the human eye to visually survey an outdoor scene, this task is highly problematic for a computer. In computer vision, the image is divided into elements (pixels). Each pixel carries information about color and luminance (defined here as intensity information). To pinpoint a certain part of an image, the computer must use meaningful selection criteria, such as object appearances. The problem is complicated by the fact that people dress in colors that sometimes blend with the background; wear hats or carry bags; and stand, walk, and change direction unpredictably. Further, the appearance of the background environment varies considerably with the presence of stationary elements, such as buildings, street signs, traffic signals, and parked cars, as well as moving objects, such as moving vehicles, bicycles, and pedestrians.

The traditional approach to computer-aided pedestrian detection is based on the use of a single video camera. Three different methods are in use to date: template matching, background subtraction, and motion-based detection.

Intelligent Vehicle Initiative

FHWA's work in multipedestrian tracking is one of the latest studies under the U.S. Department of Transportation's (USDOT) Intelligent Vehicle Initiative (IVI), a research and development program focused on vehicle safety and driver information systems. IVI is a multiagency effort involving FHWA, the National Highway Traffic Safety Administration, and the Federal Transit Administration. USDOT's Intelligent Transportation Systems (ITS) Joint Program Office provides a single budget for the agencies' vehicle-related ITS projects, economizing resources and fostering synergies in research. IVI's role is to research and advance integrated concepts that have safety implications or benefits that are not likely to be accounted for by the marketplace or are too far from commercialization to be of interest to most companies.

IVI projects have investigated the human factors, user acceptance, and technical development of individual driver information systems, advanced collision-avoidance and vehicle safety systems, and automated highway systems. An increasing number of IVI applications rely on pedestrian monitoring: traffic control, security monitoring, pedestrian flow analysis, and pedestrian counting, among others.

In the *template matching* approach, researchers create a library of possible visual patterns of pedestrians to seek similarity between a segment of the actual video frame and a library image, or *template*. Once such similarity is found, the frame part is classified as a pedestrian image. Although this approach can be useful in some circumstances, the FHWA researchers deemed this approach not very efficient because it requires the creation of a huge library of templates.

The goal of the *background subtraction* approach is to extract the pedestrians from the background. The key here is prior knowledge of the background setting. If the background is fixed, the researcher can easily "subtract" the pedestrian from subsequent registered images. When the background is not fixed, certain statistical background models may be substituted. The major drawback of this approach is the lack of generalization, which means the model may not be valid when the background settings are relatively new.

In *motion-based detection*, the researcher assumes that pedestrians are moving and that all moving objects can be treated as potential pedestrians. Although it is relatively easy to detect moving objects from registered image frames, researchers have found that it is not so easy to discriminate moving pedestrians from other moving objects, such as vehicles.

These traditional approaches may be insufficient for real-world conditions with multiple people and moving backgrounds. A recently proposed three-dimensional (3-D) stereovision approach allows distances to features to be estimated, thus enabling identification and tracking of pedestrians. This approach uses two video cameras surveying the same area from different viewpoints. The difference in the viewpoints causes a relative displacement, or disparity, of the

corresponding features in the stereo images. This disparity prompts the system to embed information on distance between objects and the cameras. Typically, when a single monocular camera captures a 3-D scene, this information on distance is lost. With stereovision, the distance to a vertically aligned object such as a pedestrian will be relatively shorter than that to the ground beyond the pedestrian or to background objects around the pedestrian. This is the key feature in pedestrian detection using stereovision.

Even though use of a stereo image is a major step forward in finding vertically aligned objects among background "noise," several challenges still need to be overcome to apply the method to pedestrian detection and tracking.

Why Use Stereo Disparity?

Stereopsis is the process in visual perception leading to depth perception, or the distance of an object from a viewer. The term comes from two Greek roots: *stereo*, meaning solidity, and *opsis*, meaning vision or sight. Computer stereovision entails recovering the 3-D information from two images of the same scene taken by two cameras in slightly different locations. Motion analysis looks at a sequence of images taken at different times and attempts to locate and measure movement between them. The challenges of using stereopsis and motion analysis are similar, both essentially involving a correspondence problem, a process of matching the same object in two or more images. In both cases the key task is to locate the image of a scene point in a set of images.

To create computer stereopsis requires three conditions. First, the stereo cameras each need to have a pair of thin lenses. Next, the two focal rays of both lenses have to be parallel and perpendicular to the stereo baseline. And, finally, the image planes of both lenses are co-linear, which implies that both of these planes lie on a single plane. Assuming that these criteria are met implies that the third axis, *z*, can be omitted in a disparity analysis. By properly rotating the *x*-*y*-*z* coordinate system, researchers can visualize a 3-D formation in a 2-D representation.

One of the main advantages of using a stereovision system is to relate the distance between the object and camera baseline and the disparity obtained from two images taken by the right and left lenses. When the object lies outside the focal region (a rectangular region between two lenses), the disparity is essentially the difference between the locations of a scene point in both right and left images. Many researchers have concluded that range is inversely proportional to the disparity. However, this statement is correct *only* when the scene point is located outside the focal region.

When the object point is inside the focal region, the range and disparity are not inversely proportional. To minimize the impact of the discrepancy, the baseline of the stereo camera needs to be small. A small baseline also facilitates estimating the disparity values (or "disparity map" when the entire image is considered) because both right and left lenses will most likely take images from the same scene.

(Top) This stereo camera is attached to a traffic signal pole at an intersection. ***Photo: Migma Systems, Inc.***

(Bottom) FHWA researchers installed a stereo camera, mounted on a traffic signal pole at this intersection in Norwood, MA. ***Photo: Migma Systems, Inc.***

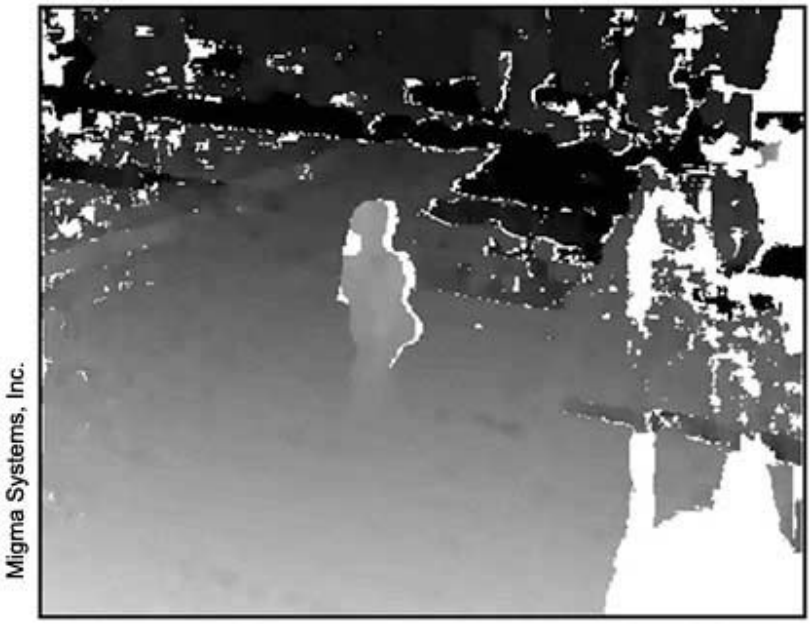
Now the remaining question is how to estimate the disparity values. In a typical disparity estimation scheme, finding conjugate points in the right and left images is one of the main problems. The search is typically based on a matching process that estimates the similarity of points in the two images on the basis of local or punctual information. Researchers typically use one of three methods: the correlation- or feature-based approach, intensity-based approach, or phase-based approach.

Feature-based methods require fairly advanced image analysis to identify features by treating pixels differently. They can be more reliable than intensity-based methods, especially for long-range correspondences, if a sufficiently dense and reliable feature map can be computed. Feature-based approaches also suffer from high computational load and classification problems.

Intensity-based methods do not require feature identification, as all pixels are treated identically. Apart from block matching, intensity-based methods using deterministic or statistical methods to solve the matching problem of disparity estimation have been predominant.

Over the past decade, many researchers have viewed the *phase-based approach* as one of the most effective methods of disparity estimation. This approach relies on the fact that the depth information from the left and right raw images can be related to the local phase difference between them.

In the FHWA stereovision system, the disparity is estimated based on feature matching.



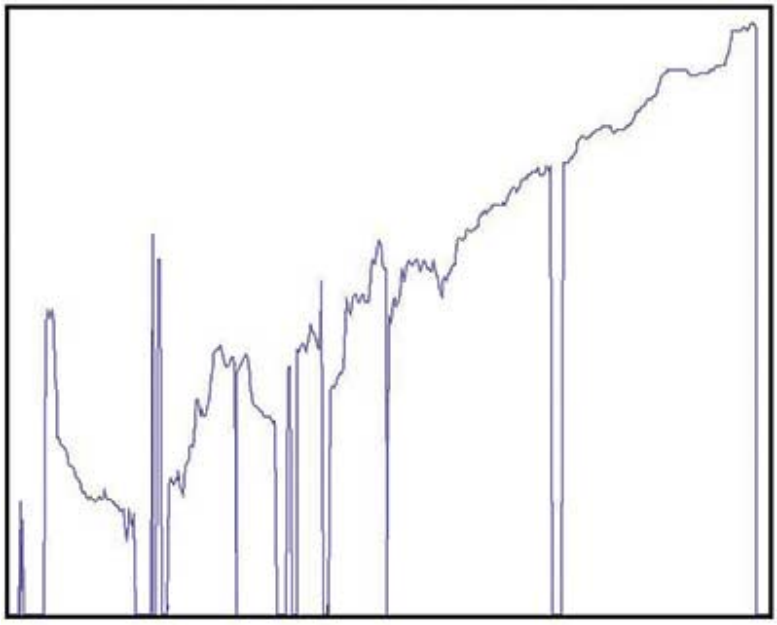
This disparity map shows a pedestrian in the center. As sometimes happens in disparity maps, the pedestrian's legs have merged into the background. The holes (white patches) represent the invalid disparity values, which mean that no disparity information at those pixels is available.

Stereovision System at FHWA

The FHWA stereovision system features a camera that has two lenses (left and right) with a resolution of 1,024 by 768 pixels, a 12-centimeter (4.7-inch) baseline, and a 70-degree horizontal field of view. As the camera is not self-powered, the researchers connected it to a laptop, which provides the power to the stereo camera. The researchers used the application programming interface from the camera's manufacturer for data collection and estimating the disparity map. Various stereo parameters, such as image resolution, mask size, validation, and disparity range, were used for calculating the disparity. The researchers used a software program to adjust the

values of the parameters for the images. To collect data, the camera was mounted on top of a pole 2.7 to 7.3 meters (9 to 24 feet) high. The pole was tied to a traffic light pole at an intersection.

In the FHWA system, the data sampling rate was 5 to 15 frames per second (fps). The actual sampling time depends on the speed of both data transmission and the computing platform. The researchers chose a sampling time of 5 fps. After analyzing the average speed of a pedestrian walking across a street, they decided to set the data collection time at 12 seconds, which is equivalent to 60 frames per dataset. The researchers then used the stereovision system to record pedestrians crossing street intersections. A total of six sets of images were collected during sunny, partly cloudy, and cloudy days. Instances when both single and multiple pedestrians were crossing the intersection were included for each weather condition. (Note: The datasets are available from FHWA upon request. See the authors' note at the end of the article.)



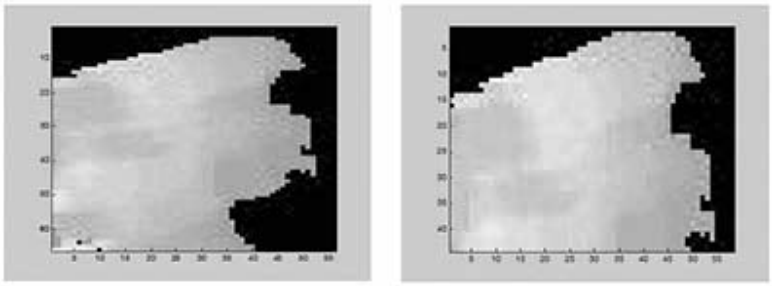
In this figure the y axis represents the disparity values and the x axis represents the column coordinates in the disparity map. The researchers observed that the disparity values increase (although not monotonically). Therefore they can detect the pedestrians in individual layers of the disparity map. ***Source: Migma Systems, Inc.***

Disparity Layered Thresholding

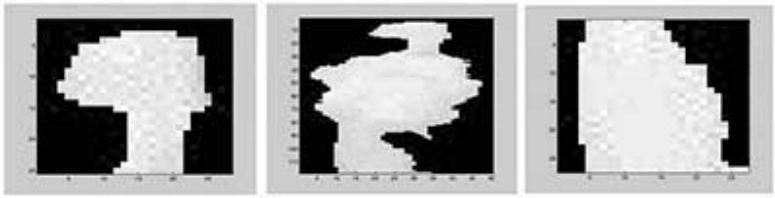
A disparity map is often used in stereovision systems. In general, the size of the map is the same as the size of either the right or left image. Each pixel represents the disparity value at the corresponding pixel, not the intensity value of the image. Since the disparity value is somewhat related to the range (distance between the object and stereo camera baseline), a disparity map can be viewed as a 3-D image. Holes or white patches in the map represent the invalid disparity values, which mean there is no disparity information at those pixels.

In a disparity map, some elements of the background near some features of the pedestrian, such as the pavement and the pedestrian's feet, are at approximately the same distance from the camera. This means that the disparity values will be similar and that it is difficult to separate the pixels associated with the pedestrian's feet from those associated with the pavement by the disparity values alone. The shapes associated with the pedestrian in the map may vary from those in the original right or left images. In fact, the pedestrian shapes in the map largely depend on the way disparities are calculated, which implies that the texture of the clothes a pedestrian is wearing will alter his or her disparity shapes. Therefore, the traditional template-matching approach will be less reliable in detecting the pedestrian in a disparity map.

To overcome this problem, the researchers developed a "thresholding" method based on the estimation of background boundary values. Because the disparity values are usually (but not always) inversely proportional to the ranges, they expected that the disparity values along the vertical lines would increase in a disparity map. In such a plot, the *y*-axis represents the disparity values and the *x*-axis represents the horizontal coordinates in the disparity map. The researchers observed that the disparity values increased (although not monotonically due to the noise in the disparity map), therefore they could analyze the disparity map in a layered fashion and detect the pedestrians in each layer.



This figure shows two small disparity chips representing various background parts. *Source: Migma Systems, Inc.*



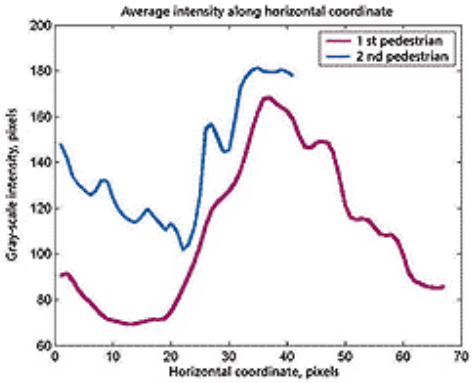
This figure shows three small disparity chips representing various body parts of a pedestrian. *Source: Migma Systems, Inc.*

To separate the pedestrians from the background, the researchers developed an estimation algorithm for the background boundary. They estimated the values for the boundary in each thresholded disparity map. For each map, they applied the morphological operator, a method used in image processing to analyze structured objects, to fill small holes and convert the resulting disparity layer into a binary image. In the binary image, they used a linear regression model to construct a background boundary line that was used to window out all potential pedestrians in the layered disparity map. With this approach, the potential pedestrian was

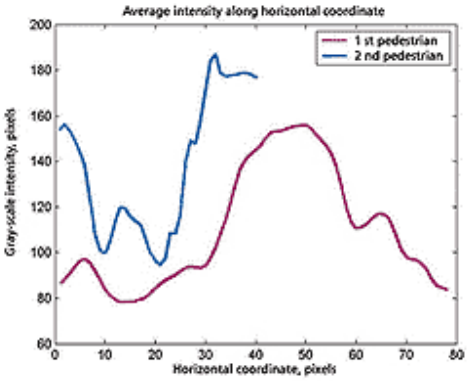
gradually extracted from the image. The number of disparity layers depended on the original map. In other words, the researchers did not fix the number of disparity layers.

Frame k with two windowed pedestrians.

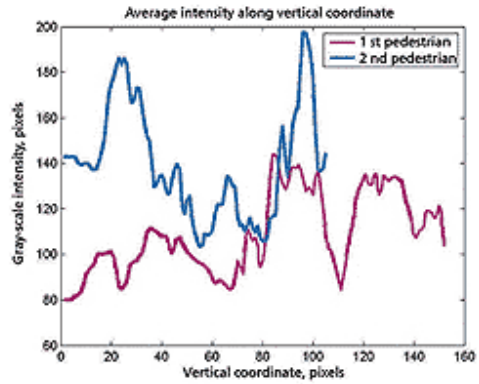
Frame k_2 with two windowed pedestrians.



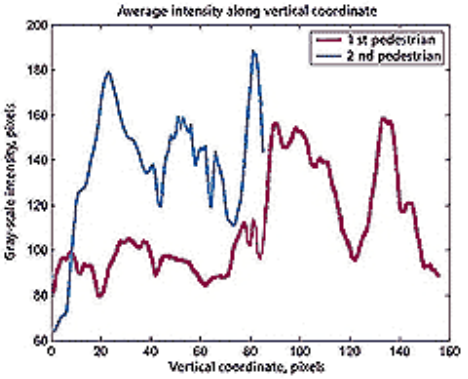
X-projection intensities for the selected pedestrians, frame k_1.



X-projection intensities for the selected pedestrians, frame k_2



Y-projection intensities for the selected pedestrians, frame k_1.



Y-projection intensities for the selected pedestrians, frame k_2.

This figure illustrates the principle of continuous pedestrian tracking. The two frames of video image (top left, k_1, and top right, k_2) were shot with a time difference of 1.2 seconds. Red and blue windows indicate

corresponding pedestrians in the two frames. Both the pedestrian images and the corresponding windows appear different in these frames. The projection intensity curves, however, shown in the plots below, remain quite similar, which makes it possible to use the projection intensity as a distinctive pattern feature. In addition, the pixel distance between the same pedestrians on the different frames is usually smaller than the distance between different pedestrians, though this may not always be true. The simultaneous use of these two features permits both continuous tracking of the detected pedestrians and elimination of background objects that were accidentally classified as pedestrians. Source: Migma Systems, Inc.

Detecting Pedestrians Using 3-D Convex Curvature

A common problem associated with the template-matching approach is the selection of a template library general enough to account for the changes in rotation, scale, brightness, and viewpoints found in the real world. To overcome this problem, the researchers developed a new method for pedestrian detection based on the disparity map. Instead of applying existing templates, they used a 3-D convex curvature feature to detect pedestrians in each layer of the map.

This figure shows two consecutive images (left, top and bottom), and their respective disparity maps (right, top and bottom), taken at a sampling time of 200 milliseconds.

The researchers developed an algorithm to extract "chips," or parts, of the pedestrian's body (head or upper body, for example) and the background at various disparity layers. Without being able to match the templates with the various body parts, it is almost impossible to know whether a particular chip contains only the head or head and upper body. As noted earlier, pedestrians in a disparity map do not have regular shapes. In other words, the researchers concluded that it would be impossible to build a library of templates for all possible shapes of pedestrian body

parts, and therefore they decided that a different approach would have to be taken to detect the pedestrian in the disparity map.

The approach the FHWA researchers took was to extract 3-D features. In particular, they looked for curvatures in the disparity chips. Unlike street signs and other background objects in the roadway environment, the human body has distinctive 3-D curvatures similar to a sphere (head), plane (body), and cylinder (legs). The researchers therefore used the 3-D convex feature to detect the pedestrian in a disparity chip. In particular, for each row of chips, a parabolic curve was used to fit the row's disparity values. The researchers then checked whether the fitting curve was convex or concave. They applied this discrimination rule to the disparity chip extracted at each layer. If a pedestrian body part (such as the head or upper body) was detected, it could be extracted out, with the window size estimate based on the chip size. Once the entire frame was processed, all the small windows were combined to form a large window that contains the pedestrian.

Continuous Pedestrian Tracking

The detection method described above neither guarantees that the pedestrians will be detected in the same order in each video frame nor that the resulting analysis will be free of false alarms (false detection). The ultimate goal of the entire system is to continuously track pedestrians and predict their location. The researchers therefore had to ensure continuous tracking of detected pedestrians and eliminate false alarms. In identifying the pedestrians and discerning their images from false alarms, the research team used two independent features: the geometric location and the average projection intensities of the images along either the rows (x intensity) or columns (y intensity).

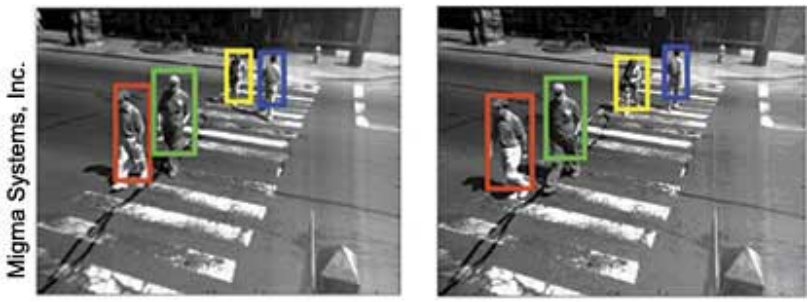
Many researchers have used the geometric location of the image window center on the image frame in pattern recognition. Others have developed and implemented studies using projection intensities.

System Test Results

The researchers tested their pedestrian detection method with a large number of images taken with the stereovision system. The images were captured at actual street intersections during sunny, partly cloudy, and cloudy days. Both single and multiple pedestrians were recorded.

The researchers then examined and analyzed the images and their respective disparity maps. For example, they looked at two images that were taken at an intersection on a sunny day, each showing four pedestrians crossing the street in a crosswalk. From the disparity maps they found that it was easy to identify the first two pedestrians, those closest to the camera. The other two pedestrians, although visible in the disparity maps, are not so clear. The researchers expected this to be the case because the latter pedestrians were about 6.4 meters (21 feet) away from the camera, while the former pedestrians are about 2.7 meters (9 feet) from the camera.

To detect the four pedestrians, the researchers first used the disparity thresholding algorithms to obtain a set of disparity layers. They then detected the potential pedestrians in each layer by checking the 3-D convex curvature features. If they detected a pedestrian in a particular layer, he or she was windowed out, signifying detection. At the end, all windows were combined to extract the entire pedestrian. The researchers then applied the Bayesian classifier, a statistical method for object classification, to associate the pedestrians detected over consecutive image frames. The approach successfully detected all four pedestrians, and the researchers windowed each with a different color coding to facilitate correlating the pedestrian association in consecutive frames.



The approach successfully detected all four pedestrians. Each is windowed with a different color coding, indicating the pedestrian association in consecutive frames.

Conclusion

In summary, the new multipedestrian detection system comprises several key modules: (1) a disparity-layered thresholding method that can be used to extract potential pedestrians layer by layer, (2) pedestrian detection with a 3-D convex curvature feature that can be used to window out pedestrians in each disparity layer, (3) continuous pedestrian association that correlates the same pedestrian in subsequent frames, and (4) estimation of pedestrian speed and location.

The initial focus of the data collection for this research was limited to collecting data on pedestrians in the crosswalk in order to validate the research approach. Phase II will extend this to detecting and tracking pedestrians both in the crosswalk and on the curb.

After testing the system using actual images taken with a stereovision system at street intersections, the research team concluded that the results show that the system has significant potential for use in Intelligent Vehicle Initiative (IVI) applications.

Eventually, the researchers will incorporate the software algorithms into selected hardware platforms deployed at intersections, using wireless communications to traffic controllers. FHWA also needs to further improve the technologies to increase the detection accuracy and reduce the system cost.

Once the detection system developed in this research is refined, State departments of transportation will be able to use it in their pedestrian safety programs. As South Carolina's Dodds says, "Systems that better enable pedestrians and motorists to share the pavement safely at the same time would be worthy of consideration."

David R. P. Gibson is a highway research engineer on the Enabling Technologies Team in FHWA's Office of Operations Research and Development. He is a registered professional traffic engineer with bachelor's and master's degrees in civil engineering from Virginia Polytechnic Institute and State University. His areas of interest include traffic sensor technology, traffic control hardware, traffic modeling, and traffic engineering education.

Bo Ling received his M.S. in applied mathematics in 1990 and a Ph.D. in electrical engineering in 1993 from Michigan State University. Ling has served as a principal investigator for numerous government-funded research projects. He is a cofounder and president and chief executive officer of Migma Systems, Inc. He became a senior member of the Institute of Electrical and Electronics Engineers, Inc., in 1998 and is a part-time faculty member of the Department of Electrical and Computer Engineering at Northeastern University in Boston.

Michael Zeifman has a degree in materials science and metallurgy from Leningrad Polytechnic Institute (now St. Petersburg State Polytechnic University) in Russia (1987), an M.Sc. in quality assurance from Technion-Israel Institute of Technology in Israel (1997), and a Ph.D. in physical reliability, also from Technion-Israel Institute of Technology (2000). After years of industry and university research, he joined Migma Systems, Inc., as a senior research engineer in 2005.

Shaoqiang Dong received his Ph.D. from the University of Massachusetts at Amherst in 2004. He received his M.S. (1997) and B.S. (1994) degrees, both in mechanical engineering, from Xi'an Jiaotong University in China. He joined Migma Systems, Inc., as a research engineer in 2004.

Uma Venkataraman received her M.S. in computer science from Madras University in India (1996), and she has several years of software development experience. She joined Migma Systems, Inc., as a senior software engineer in 2003. She focuses on software development.