PAPER

Layered Detection for Multiple Overlapping Objects

Hironobu FUJIYOSHI[†a)] ***and*** **Takeo KANADE**[††,†††]**,** ***Members***

SUMMARY This paper describes a method for detecting multiple overlapping objects from a real-time video stream. Layered detection is based on two processes: pixel analysis and region analysis. Pixel analysis determines whether a pixel is stationary or transient by observing its intensity over time. Region analysis detects stationary regions of stationary pixels corresponding to stopped objects. These regions are registered as layers on the background image, and thus new moving objects passing through these layers can be detected. An important aspect of this work derives from the observation that legitimately moving objects in a scene tend to cause much faster intensity transitions than changes due to lighting, meteorological, and diurnal effects. The resulting system robustly detects objects at an outdoor surveillance site. For 8 hours of video evaluation, a detection rate of 92% was measured, which is higher than traditional background subtraction methods.

key words: *object detection, video surveillance, activity recognition*

1. Introduction

Recently, automated video surveillance using video understanding technology has become an important research topic in the area of computer vision [1]. Within video understanding technology for surveillance use, detection of moving objects in video streams is known to be a significant, and difficult, research problem [2]. Conventional approaches to moving object detection include temporal differencing [3], [4], background subtraction [2], [5]–[7], and optical flow [8]–[10]. One of the most successful approaches to date is adaptive background subtraction [6]. The basic idea is to maintain a running statistical average of the intensity at each pixel – when the value of a pixel in a new image differs significantly from this, the pixel is flagged as potentially containing a moving object. One problem of this approach, along with other conventional approaches to motion detection is that objects that cease moving within the image simply disappear from the representation. A robust detection system should continue to "see" objects that have stopped and disambiguate between overlapping objects in the image. For example, a car that comes into the scene and parks should not be considered as part of the scene background. However, its stationary pixels should play the role of background for detecting the motion of a person getting out of the car.

We have developed a novel approach to object detection based on layered adaptive background subtraction. Layered detection is based on two processes: pixel analysis and region analysis. Pixel analysis determines whether a pixel is stationary or transient by observing its intensity value transitions over time. The technique is derived from the observation that moving objects under observation cause much faster intensity transitions than changes due to lighting or weather. Region analysis outputs a stationary region consisting of stationary pixels as a stopped object. This region is registered as a layer on the background image, allowing new moving objects passing through the layer to be detected.

The paper is organized as follows. In Sect. 2, we describe the algorithm of layered detection based on two processes. In Sect. 3, we describe evaluation method based on time duration and experimental results of 8 hours evaluation, then show the effectiveness of the layered method.

2. Layered Detection Algorithm

Layered detection is based on two processes: pixel analysis and region analysis. The purpose of pixel analysis is to determine whether a pixel is *stationary* or *transient* by observing its intensity value over time. Region analysis deals with the agglomeration of groups of pixels into moving regions and stopped regions. Figure 1 graphically depicts the process. By observing the intensity transitions of a pixel, dif-

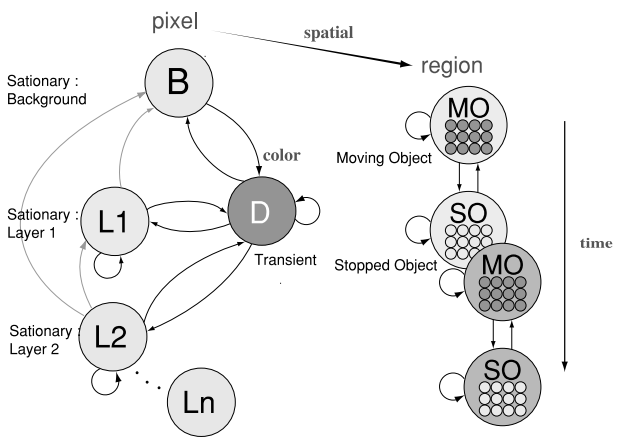

Fig. 1 The concept — combining pixel statistics with region analysis to provide a layered approach to motion detection.

Manuscript received December 14, 2003.
Manuscript revised May 22, 2004.

[†]The author is with the Department of Computer Science, Chubu University, Kasugai-shi, 487–8501 Japan.
[††]The author is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA.
[†††]The author is with the Digital Human Research Center, Advanced Industrial Science and Technology (AIST), Tokyo, 135–0064 Japan.
a) E-mail: hf@cs.chubu.ac.jp

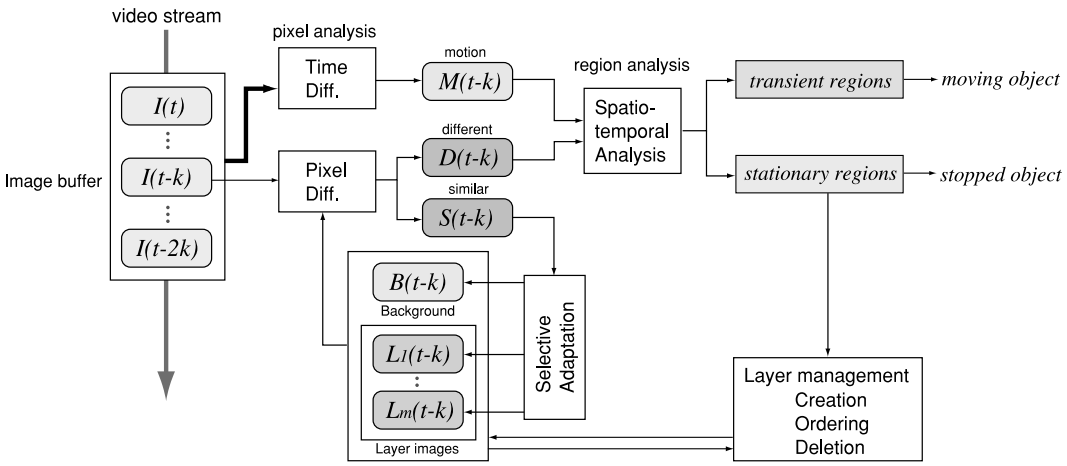

Fig. 2 Architecture of the detection process. Temporal analysis is used on a per pixel basis to determine whether pixels are transient or stationary. Transient pixels are clustered into groups and assigned to spatio-temporal layers. A layer management process keeps track of the various background layers.

ferent intensity layers, connected by transient periods, can be postulated. Within each pixel, an intensity state is started from the background value. When a moving object passes through, the intensity state changes to transient. When the object stops, the state reverts to a stationary value, which may differ from the value of the background. Therefore, it is clear that moving objects and stopped objects can be detected by keeping track of collections of transient and stationary pixels.

Figure 2 shows the architecture of the detection processes. A key element of this algorithm is that it needs to observe the behavior of a pixel for some time before determining if that pixel is undergoing a transition.

2.1 Pixel Analysis

In an outdoor surveillance scenario, it has been observed that a pixel's intensity value displays three characteristic profiles depending on what is occurring in the scene at that pixel location.

- An object moving through the pixel displays a profile that exhibits a step change in intensity, followed by a period of instability, then another step back to the original background intensity. Figure 3 (a) shows this profile.
- An object moving through the pixel and stopping displays a profile that exhibits a step change in intensity, followed by a period of instability, then it settles to a new intensity as the object stops. Figure 3 (b) shows this profile.
- Changes in intensity caused by lighting or meteorological effects tend to be smooth changes that don't exhibit large steps. Figure 3 (c) shows this profile.

To capture the nature of changes in pixel intensity profiles, two factors are important: the existence of a significant step change in intensity, and the intensity value to which the profile stabilizes after passing through a period of instability. To interpret the meaning of a step change (e.g. object passing through, stopping at, or leaving the pixel), we need to observe the intensity curve re-stabilizing after the step change. This introduces a time-delay into the process. In particular, current decisions are made about pixel events k frames in the past. In our implementation k is set to correspond to one second of video. Therefore, although our algorithms runs in "real-time", there is a lag of one second. This delay is more than made up for by the improved quality of detections, resulting from having knowledge of the future when making decisions about the "current" frame.

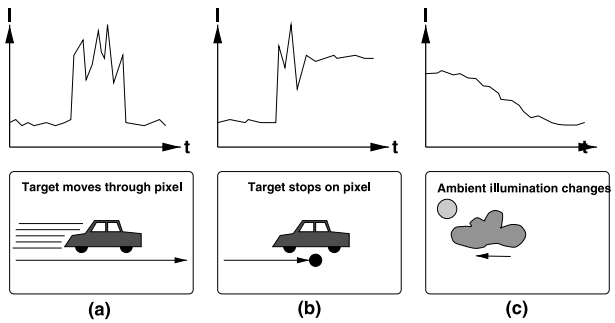

Fig. 3 Characteristic pixel intensity profiles for common events. Moving objects passing through a pixel cause an intensity profile step change, followed by a period of instability. If the object passes through the pixel (a), the intensity returns to normal. If the object stops (b), the intensity settles to a new value. Variations in ambient lighting (c) exhibit smooth intensity changes with no large steps.

Let I_t be some pixel's intensity at a time t occurring k frames in the past. Two functions are computed: a motion trigger T just prior to the frame of interest t, and a stability measure S computed over the k frames from time t to the present. The motion trigger is simply the maximum absolute difference between the pixel's intensity I_t and its value in the previous five frames:

$$T = \max \left\{ |I_t - I_{(t-j)}| , \forall j \in [1,5] \right\} \quad (1)$$

The stability measure is the variance of the intensity profile from time t to the present:

$$S = \frac{k\sum_{j=0}^{k} I_{(t+j)}{}^2 - (\sum_{j=0}^{k} I_{(t+j)})^2}{k(k-1)} \quad (2)$$

At this point a transience map M can be defined by the following algorithm for each pixel, taking three possible values: background=bg; transient=tr and stationary=st. Background intensity is prepared in advance as a background image.

```
if ((M = st or bg) AND (T > Threshold))
   M = tr
}
if ((M = tr) AND (S < Threshold)) {
   if (I = background intensity)
      M = bg
   else
      M = st
}
```

Background is updated by an Infinite Impulse Response (running average) filter to accommodate slow lighting changes and noise in the imagery, as well as to compute statistically significant step-change thresholds [12].

$$B(t) = \alpha I(t) + (1-\alpha)B(t-1) \quad (3)$$

The constant α determines how fast the background is allowed to change.

2.2 Region Analysis

Non-background pixels in the transience map M are clustered into regions R_i using a nearest neighbor spatial filter with clustering radius r_c. This process is similar to performing a connected components segmentation. However, gaps up to a distance of r_c pixels can be tolerated within a component. Choice of r_c depends upon the scale of the objects being tracked. A clustered region has one of the following three states:

- All pixels in region are labeled as transient. The region must be a moving object.
- All pixels in region are labeled as stationary. The region must be a stopped object.
- The region contains a mixture of transient and stationary pixels. The region may contain both stopped and moving objects.

Each spatial region R is then analyzed according to the following algorithm:

```
if (R = tr) {
    R -> moving object
}
elseif (R = st) {
    %remove all pixels already assigned
    %to any layer
    R = R - (L(0) + L(1) + .. + L(j))
    %if anything is left, make new layer
    if (R != 0) {
        make new layer L(j+1) = R
        R -> stopped object
    }
else {
    %R contains a mixture of tr and st
    R = R - (L(0) + L(1) + .. + L(j)
    SR(i) = spatial_clustering(R)
    for (each region SR(i)) {
        if (SR = tr) {
            SR -> moving object
        }
        if (SR = st) {
            make new layer L(j+1) = SR
            SR -> stopped object
        }
        if (SR = (st + tr)) {
            SR -> moving object
        }
    }
}
```

Where L(j) is a layer image and j is the number of layer images that are already registered.

Regions that consist of stationary pixels are added as a new layer over the background. A layer management process is used to determine when stopped objects resume motion or are occluded by other moving or stationary objects. When an object that is already registered as a layer starts to move, the layer is deleted by the layer manager. Intensity values within stationary layered regions are updated by an IIR filter in the same way that the background is updated.

3. Detection Results

3.1 Example of Analysis

Figure 4 shows an example of the analysis that occurs at a single pixel. The video sequence contains the following activities at the pixel:

1. A vehicle drives through the pixel and stops
2. A second vehicle occludes the first and stops
3. A person, getting out of the second vehicle, occludes the pixel
4. The same person, returning to the vehicle, occludes the pixel again
5. The second car drives away
6. The first car drives away

As can be seen, each of these steps is clearly visible in the pixel's intensity profile, and the algorithm correctly identifies the layers that accumulate.

Figure 5 shows the output of the region-level layered detection algorithm. The detected regions are shown sur-

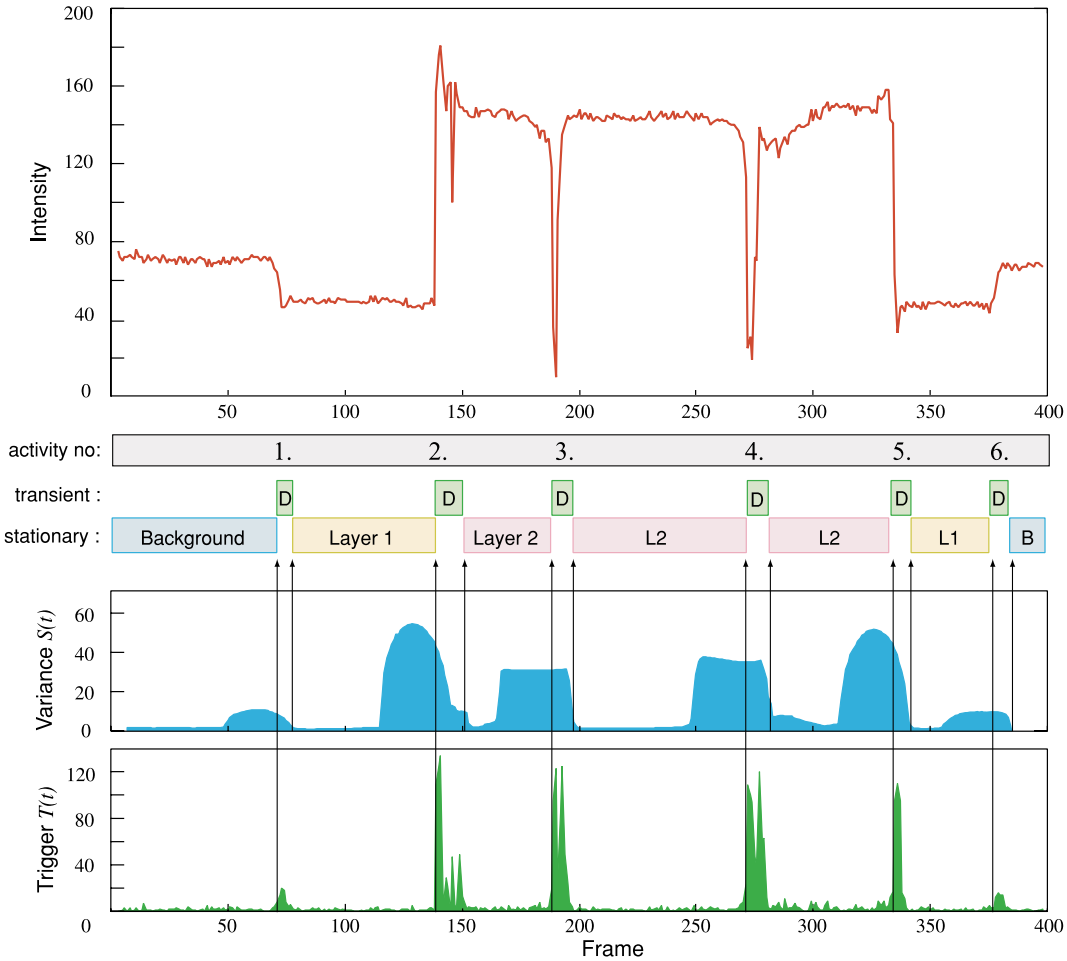

Fig. 4 Example pixel analysis of the scene shown in Fig. 5. A car drives in and stops. Then a second car stops in front of the first. A person gets out and then returns again. The second car drives away, followed shortly by the first car.

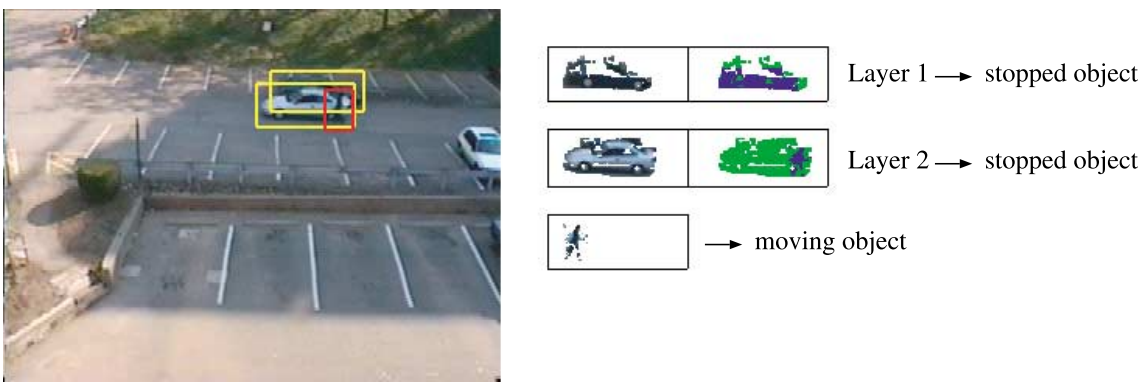

Fig. 5 Detection result. Here one stopped vehicle partially occludes another, while a person in moving in the foreground. Displayed on the right are the layers corresponding to the stopped vehicles and the moving foreground person, together with bitmaps denoting which pixels are occluded in each layer.

rounded by bounding boxes — note that all three overlapping objects are independently detected. Each stopped car is depicted as a temporary background layer, and the person is determined to be a moving foreground region overlayed on them. The pixels belonging to each car and to the person are well disambiguated.

3.2 Evaluation Method

To measure the performance of a detection algorithm, a

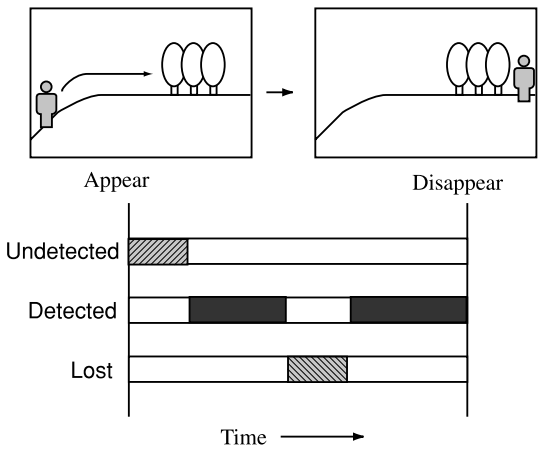

Fig. 6 Evaluation.

number of criteria are relevant. An object should be detect as soon as possible after it appears. Once the object is first detected, it should be continuously tracked until it disappears from view. Occasionally a detected object will become lost due to low contrast or other factors as it passes through the field view. We propose an evaluation method based on measuring time durations (Fig. 6), specifically, the amount of time an object remains undetected, the amount of time it remains detected (tracked), and the amount of time it is lost during its traversal of the field of view. A human operator evaluates video footage to determine these time durations. The detection rate and the loss rate are then calculated as:

$$Detection = \frac{\sum D}{\sum U + \sum D + \sum L} \quad (4)$$

$$Loss = \frac{\sum L}{\sum D + \sum L} \quad (5)$$

U : Undetected time duration [s]
D : Detected time duration [s]
L : Loss time duration [s]

In this definition, "U: Undetected" means the undetected time duration after an object appears until the object is detected. Once the object is detected, an undetected time duration is evaluated as loss time "L:Loss".

Using this evaluation method, the real peformance of detection is computed. Note that false positives are not considered in this evaluation. In case of false positives, we just count a total number of false positives.

3.3 Experimental Results

On an Pentium III running at 500 MHz, our method can process 6 to 9 frames a second (frame size 320×240 pixels). The variation in the frame rate is due to the size and amount of moving objects. This detection algorithm has been evaluated on eight hours of video tape for which ground-truth labeling of moving obects (people and vehicles) was manually

Table 1 Detection rates (loss rate) [%].

	MTD		Layered method	
	camera1	camera2	camera1	camera2
Sunny day	58.4 (29.1)	89.9 (11.2)	84.4 (10.4)	94.7 (5.5)
Cloudy day	92.1 (8.5)	93.8 (6.5)	92.7 (7.8)	96.2 (3.8)
Average	83.5		92.0	
False positive	18 times		20 times	

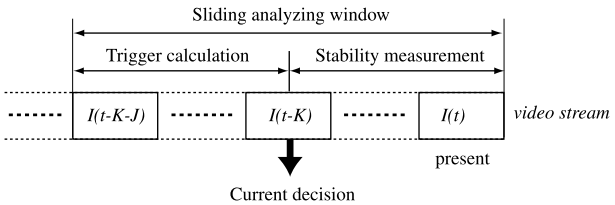

Fig. 7 Sliding analysing window.

determined. Four hours of data were taken on a sunny day, and four hours on a cloudy day. Probability of detection was determined by time duration evaluation as described above. Sometimes, there are objects which are not detected totally because of the low contrast. The loss rate does not contain the case mentioned above in order to evaluate the undetected time after the object is detected. Therefore, the sum of the detection rate and the loss rate sometimes exceeds 100%, because the denominators of Eqs. (4) and (5) are different.

Table 1 shows detection rates by MTD and layered detection. MTD is the standard adaptive background subtraction method described in [11]. Note that MTD does not have the capability to continue detecting stopped objects. Failures of MTD and layered method under sunny conditions are mainly due to loss of contrast in shadowed areas. This is one reason why cloudy day perfomance is better for both cameras.

On the other hand, in this situation, our method can detect additional objects that MTD cannot because the threshold value for a motion trigger T of Eq. (1) can be set to a low value for detecting a small change in intensity over time. Although we may get some false positives due to the more sensitive threshold value, our method can suppress those false positives because the method distinguishes whether the pixel is stationary or transient by analyzing the variance of the intensity profile from $t - K$ to the present as shown in Fig. 7.

There are 20 false positives with layered detection and 18 with MTD. Most of these occured on a sunny day, because there are false positives on the front windows of vehicles due to reflection of the sun. However, this is not a problem, because it is possible to eliminate these false positives at the next stage of tracking or classification [11].

3.4 Detection in Shadow Area

There are failures of layered detection under sunny conditions due to loss of low constrast in shadow area. This is

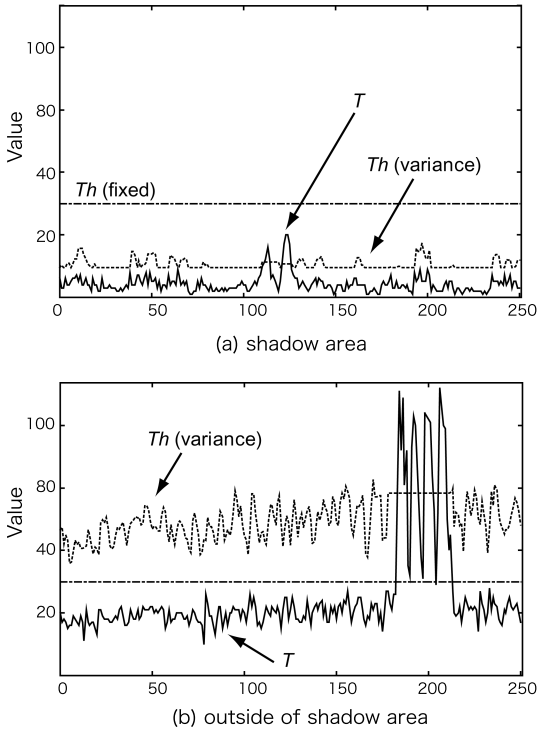

Fig. 8 Changes of motions trigger value over the time.

because the value of motion trigger T is small. In our implementation at the experiment described in Sect. 3.3, the threshold value for the motion trigger was set to constant value. If the value is set as bigger than the value of motion trigger, the motion of a object in shadow area can not be distinguished as a transient pixel. To improve detection performance in shadow area, we use adaptive thresholding based on intensity changes in the past frames. The intensity changes is calculated as a variance of intensity by the following equation:

$$S_t = \frac{K\sum_{i=1}^{k} I_{(t-i)}{}^2 - (\sum_{i=0}^{K} I_{(t-i)})^2}{K(K-1)} \quad (6)$$

The variance in the past K frames is used to set a threshold value Th_t which distinguishes whether the pixel is transient or stationary. For each pixel in the image, the threshold Th_t has to be calculated at every frame as follows:

$$Th_t = \begin{cases} 4 \cdot S_t & T_{t-1}, \leq Th_{t-1} \\ Th_{t-1} & , T_{t-1} > Th_{t-1} \end{cases}$$

Figure 8 (a) shows values of motion trigger and threshold at a pixel in shadow area and (b) shows outside of the shadow area. In shadow area, the intensity becomes small, so the motion of a object from 120th to 130th frames can not be detected by thresholding using fixed value. On the other hand, the motion can be detected by threshloding by variance, because the threshold value is chosen to adapt the intensity changes in past frames. In the case of outside of shadow area, we see that the motion of a object from 180th to 220th frames is detected, because there is no problem of low contrast.

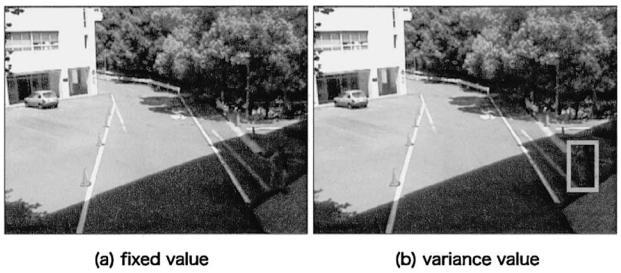

Fig. 9 Example of detection result in shadow area.

Table 2 Detection rates in shadow area [%].

MTD	Layered method	
	fixed value	variance
55.5	65.6	75.2

Figure 9 shows an example of detection result in shadow area. Table 2 shows detection results of layered method with adaptive thresholding by variance. Note that the detection rate is not high because the rate is evaluated only in shadow area. It is clear that the detection performance in shadow area is improved, because the adaptive thresholding by variance can adapt small changes in intensity at each pixels.

4. Conclusions

This paper has presented a new method for detecting regions of multiple overlapping objects from a real-time video stream. Layered detection is based on two processes: pixel analysis and region analysis. Pixel analysis determines whether a pixel is stationary or transient by observing pixel intensity over time. Region analysis outputs a stationary region consisting of stationary pixels as a stopped object. Then the region is registered as a layer on the background image. Therefore, a moving object passing through the layer can be detected as an independent object. The resulting system robustly detects objects in an outdoor surveillance site. For 8 hours of video evaluation, a detection rate of 92% was measured which is higher than traditional background subtraction methods.

Layered detection has two main advantages; it detects objects robustly because of doing a time series analysis at each pixel to detect motion, and it detects multiple overlapping objects independently by using a layer management scheme.

However, in case there are some vehicles parked in the initial background image, it might lead to wrong detection result when the vehicles move away. Choosing the initial background image is not carefully investigated yet in this paper, and this will be our next work. We will try to eliminate such wrong detection regions caused by the initial background by using texture analysis which will recognize the continuity of neighbor pixels on the edge of the mis-detected region.

Acknowledgments

The authors would like to thank the CMU VSAM team members: Robert Collins, Alan Lipton, Dave Duggins, Raju Patil, David Tolliver, Yanghai Tsin, Yong-Tae Do, Nobuyoshi Enomoto and Osamu Hasegawa for motivating this work and providing a cool working environment.

References

[1] VSAM, Section I, "Video surveillance and monitoring," Proc. DARPA Image Understanding Workshop, vol.1, pp.1–400, Nov. 1998.

[2] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," Proc. International Conference on Computer Vision, pp.255–261, 1999.

[3] C. Anderson, P. Burt, and G. van der Wal, "Change detection and tracking using pyramid transformation techniques," Proc. SPIE — Intelligent Robots and Computer Vision, vol.579, pp.72–78, 1985.

[4] P.L. Rosin and T. Ellis, "Image difference threshold strategies and shadow detection," Proc. British Machine Vision Conference, pp.347–356, 1995.

[5] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: Real-time surveillance of people and their activities," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.809–830, Aug. 2000.

[6] C. Stauffer and W.E.L. Grimson, "Learning patterns of activity using real-time tracking," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.747–757, Aug. 2000.

[7] C.R. Wren, A. Azarbayejani, T.J. Darrell, and A.P. Pentland, "Pfinder: Real-time tracking of the human body," IEEE Trans. Pattern Anal. Mach. Intell., vol.19, no.7, pp.780–785, July 1997.

[8] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," Int. J. Comput. Vis., vol.12, no.1, pp.42–77, 1994.

[9] C. Cedras and M. Shah, "Motion-based recognition: A survey," Image Vis. Comput., vol.13, no.2, pp.129–155, March 1995.

[10] G. Halevy and D. Weinshall, "Motion of disturbances: Detection and tracking of multi-body non-rigid motion," Mach. Vis. Appl., vol.11, no.3, pp.122–137, 1999.

[11] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa, "A system for video surveillance and monitoring: VSAM final report," Technical Report CMU-RI-TR-00-12, Robotics Institute, CMU, May 2000.

[12] A. Lipton, H. Fujiyoshi, and R.S. Patil. "Moving target detection and classification from real-time video," Proc. 1998 Workshop on Applications of Computer Vision, pp.8–14, Oct. 1998.

Hironobu Fujiyoshi received the Ph.D. degree in electrical engineering from Chubu University, Japan, in 1997. For his thesis, he developed a fingerprint verification method using spectrum analysis, which has been incorporated into a manufactured device sold by a Japanese security company. He is a Member of Faculty at the Department of Computer Science, Chubu University. From 1997 to 2000 he was a post-doctoral fellow at the Robotics Institute of Carnegie Mellon University, Pittsburgh, PA, USA, working on the DARPA Video Surveillance and Monitoring (VSAM) effort and the humanoid vision project for the Honda Humanoid Robot. He performs research in the areas of real-time object detection, tracking, and recognition from video.

Takeo Kanade received the B.E. degree in electrical engineering in 1968, the M.E. degree in 1970, and the Ph.D. degree in 1973 from Kyoto University, Japan. Currently, he is Helen Whitaker Professor of Computer Science at Carnegie Mellon University, Pittsburgh, PA, USA and Director of Digital Human Research Center (AIST). Dr. Kanade was a recipient of the Robotics Industry Association, Joseph F. Engelberger Award in 1995, the Japan Robotics Association, JARA Award in 1997, the Yokogawa Prize at the International Conference on Multi Sensor Fusion and Integration for Intelligent Systems in 1997, the Hip Society, Otto AuFranc Award in 1998, the Hoso Bunka Kikin Foundation Award in 1994, and the Marr Prize at The Third International Conference on Computer Vision in December 1990. He was also selected as the author of one of the most influential papers that appeared in the Artificial Intelligence journal in the last ten years in 1992. He is Founding Chief Editor of the *International Journal of Computer Vision*. He is a Member of the National Acadeny of Engineering and a Fellow of the American Association for Artificial Intelligence (AAAI).