Automatic Scene Understanding by Long-Term Observation

Takehiro Yashiro Yang Qin Ikuhisa Mitsugami Koh Kakusho Michihiko Minoh
Academic Center for Computing and Media Studies, Kyoto University
{*yashiro, yangqin, mitsugami, kakusho, minoh*}*@mm.media.kyoto-u.ac.jp*

Abstract

Opening and sharing information from the existing cameras is thought to be helpful so as to offer the new real-time and real-world oriented contents. Motivated by this idea, we have advocated the "sensing web." This paper presents a method that automatically extracts privacy-free information from images of each camera, which is one of the basic and important topic of the sensing web. This method is realized without any manual parameter tuning for each camera by utilizing the long-term observation , and so suitable to be applied to a huge number of the existing cameras.

1. Introduction

In these days, a huge number of cameras have been installed in various places in our daily living environments: stations, streets, malls, etc. Some of those cameras constitute networks for exchanging their data in order to attain the purpose for which they are installed, more efficiently. In this paper, we refer to these networks as Ubiquitous Sensor Networks (USNs). Each USN is installed by some institutions including a local government, a transit company, a security company, and so on, for some specific purposes: traffic control, building management, video surveillance, etc. The sensor data obtained from the USN is used only for the purpose by the institution exclusively. However, such sensor data can actually be used for various purposes other than their original purpose, because the data include raw real-time information of the real world. If the sensor data were opened to the public so that anyone can use the data for their own purpose, similar to the Web, the data could serve as a new worldwide social information infrastructure that supplies the information different from that supplied by current Web. In this paper, we call this new social information infrastructure the Sensing Web [1].

In realizing the Sensing Web, the privacy should be considered. This is because images obtained by the cameras may include some people so that sharing such images over the Internet may invade their privacy. For avoiding this privacy problem, we introduce the privacy filtering process that extracts privacy-free information from the images, and combine the process with every camera so as that only the filtered information is opened to the Web.

It is also very important that the process cannot use any parameters and conditions that should be manually specified according to the camera's location and scene, because it is unrealistic to specify these information to all the cameras in the world.

Considering above, this paper aims at a method that automatically extracts privacy-free information from the camera images; the numbers of humans, bikes, and cars in the observed area. This method consists of two modules. One is a new foreground extraction method based on the background subtraction method which is however robust to the sudden changes of the illumination, and the other is a method to automatically classify the foreground region into these three types of objects. Note that the method is designed not to need any parameter tuning for each camera, which is realized by utilizing the long-term observation.

2. Background Subtraction in Long-Term Observation

2.1. Background Subtraction Robust to Sudden Change of Illumination

Considering that the camera is fixed so that it observes the same place, the background subtraction is thought to be reasonable for the foreground region extraction. Denote $V(\boldsymbol{p}, t)$ as a value of a pixel $\boldsymbol{p}$ of the image captured at t. The background subtraction method calculates the difference D in the following equation:

$$D(\boldsymbol{p}, t) = V(\boldsymbol{p}, t) - V(\boldsymbol{p}, t_0). \quad (1)$$

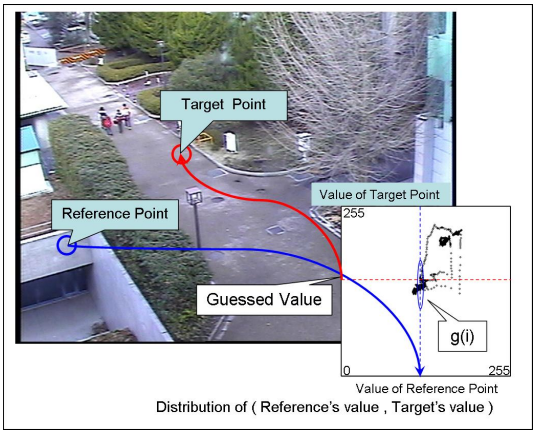
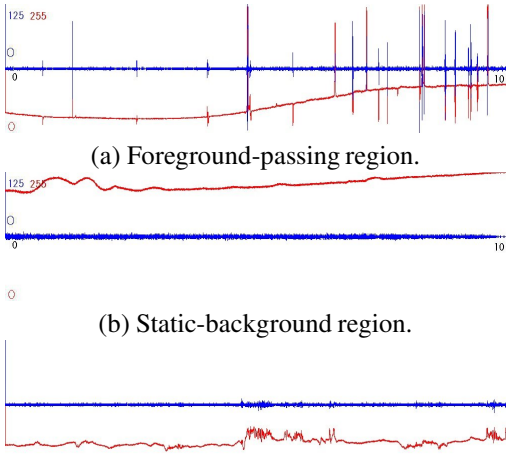


Figure 1. Background subtraction using the reference pixel.



(a) Foreground-passing region.

(b) Static-background region.

(c) Variable-background region.

Figure 2. Time variation of pixels in three regions.

In the most simple way, $V(\boldsymbol{p}, t_0)$ is chosen from an image at a past time t_0. Some sophisticated ways [2, 3, 4] use past sequential images to determine $V(\boldsymbol{p}, t_0)$. These existing ways are, however, $V(\boldsymbol{p}, t_0)$ does not contain the information of the current image, which means the image at t, so that they cannot work well when the sun light gets suddenly stronger or weaker at t.

In contrast, the proposed method introduces a reference pixel $\boldsymbol{q}$, which is defined as a pixel whose value changes most highly correlated to the target pixel $\boldsymbol{p}$ and the subtraction is calculated as follows:

$$D(\boldsymbol{p}, t) = V(\boldsymbol{p}, t) - f_{pq}(V(M(\boldsymbol{q}), t)), \qquad (2)$$

where M is a correspondence map between each $\boldsymbol{p}$ and $\boldsymbol{q}$, and f_{pq} is denoted by a relation between $V(\boldsymbol{p})$ and $V(\boldsymbol{q})$. This equation means that the background value of p is estimated not by the past value(s) of p but by the current value of q, which is expected to estimate the current value of p well. Figure 1 shows this process. Because the background value is estimated by the information of the current time t as described also in Equation (2), this method is robust to the sudden change of illumination.

2.2. How to Determine the Reference Pixel

The performance of this method is, of course, strongly depending on the reliability of the correspondence map M and the function f_{pq}. Therefore, the reference pixel q for each target pixel p should be selected properly.

As q is referred to estimate the background value of p, any foreground must not pass on q. It is also needed

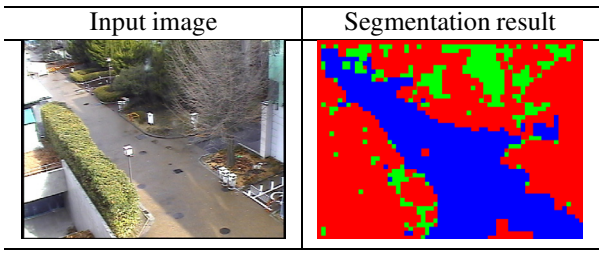


Figure 3. Segmentation result.

that q should not correspond to a nonstatic object[1] in the real world. Considering above, the image is divided into three types of regions; foreground-passing regions, static-background regions and variable-background regions. As the pixel values in these regions show different time-variation as shown in Figure 2, they can be divided using the images obtained by long-term observation. Figure 3 shows the result of this segmentation.

For each pixel on the foreground-passing region, then, the reference pixel is selected in the static-background region in order to obtain the correspondence map M. The function f_{pq} is determined by calculating the regression line of the variations of the target and reference pixel values.

2.3. Experimental Results

Figure 4 shows the results of the proposed background subtraction method. We can see that even when

[1] Trees, flags, for example, which easily swing or wave by winds.

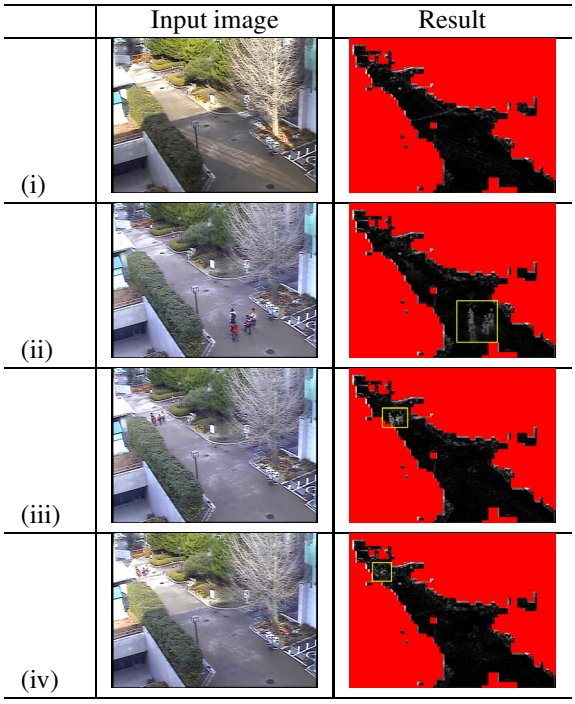


Figure 4. Background subtraction results by the proposed method.

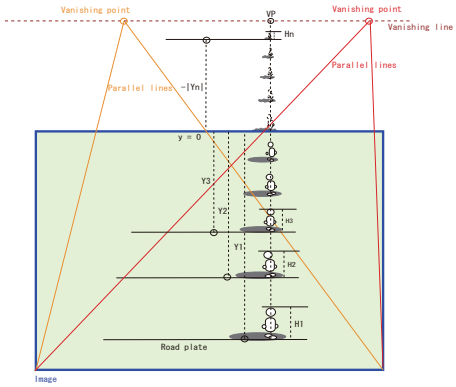


Figure 5. Influence of the perspective parameters.

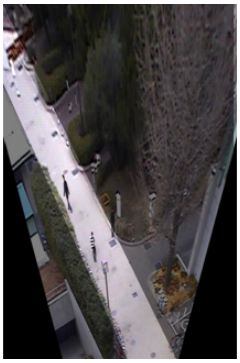


Figure 6. Evaluation of Perspective Compensation.

the sun light suddenly changes only the foreground regions are detected.

3. Object Categorization Using Mutual Difference of Size and Velocity

The traffic information is thought to be one of the useful information obtained by the surveillance cameras. To know the traffic information, we do not necessarily need the raw image of the camera but need only the number of humans, bikes, and cars in that place. Although there have been a lot of studies [5, 6] that detect and track humans and cars, most of them need some kinds of knowledge about their appearances, shapes, feature points, and so on, which have to be specified according to the position and orientation of the camera. However, in the sensing web, there are a lot of variety about the camera installation so that it is hard to construct the knowledge consistent among all the cameras.

This paper thus proposes a method that is able to automatically obtain their numbers without any knowledge, but only with the fact that relative sizes and velocities of a human, a bike, and a car in images of a camera increase in this order, which is expected to be consistent among the many cameras. To classify every foreground object into these three kinds, a classifier is generated by unsupervised learning about the collected foregrounds obtained by the long-term observation on the offline step. Once the classifier is acquired, every detected foreground is classified into these three types of object on the online step. The following sections describe the details of the offline process.

3.1. Compensation of Perspective Distortion

On using the relative sizes and velocities for the classification, the perspective distortion should be considered. Because of this distortion, a object's size and velocity on the 2D image become different according to its position, so that the classification does not work correctly. Therefore, the perspective distortion for normalizing them has to be estimated. In this paper, it is estimated by tracking multiple objects so as to collect the

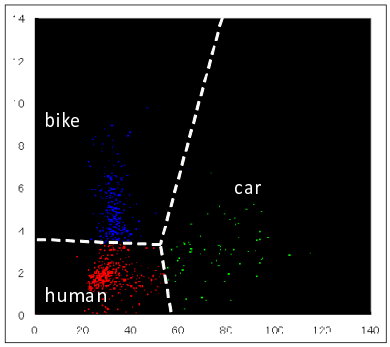


Figure 7. K-means Clustering about the object's size and velocity.

Figure 8. Results of different camera installations.

variations of their sizes and velocities. Although these objects, of course, cannot be classified on this step, the size and velocity of each of them changes similarly according to the same perspective parameters as shown in Figure 5.

Figure 6 shows the image undistorted by the estimated perspective parameters. We found the road region is well undistorted.

3.2. Compensation of Perspective Distortion

Once the perspective parameters are estimated, the detected foregrounds can be normalized. Because the sizes and velocities of humans, bikes, and cars is different statistically from one another, it is expected that these types can be classified by the k-means clustering about the size and velocity. Figure 7 shows the result of the classification.

3.3. Experimental Results

For evaluating the effectiveness of the proposed method, it is applied to some surveillance cameras at the different locations. Figure 8 shows their results. For all the cameras, the method works well so that the humans, bikes, and cars are correctly counted.

3.4. Conclusion

This paper described a new foreground extraction method based on the background subtraction method which is however robust to the sudden changes of the illumination, and a method to automatically count the number of humans, bikes, and cars captured by a camera. Each of them were implemented by utilizing the long-term observation, so as not to need any parameter tuning for each camera. The experimental results showed their effectiveness.

References

[1] M. Minoh, K. Kakusho, N. Babaguchi, T. Ajisaka "Sensing Web Project - How to handle privacy information in sensor data," 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, 2008.

[2] B.Lo and S.Velastin: "Automatic congestion detection system for underground platforms," International Symposium on Intelligent Multimedia, Video and Speech Proceeding, pp.158–161, 2001.

[3] I.Haritaoglu, D.Harwood, and L.Davis: "W4: Real-Time Surveillance of People and Their Activities," PAMI, Vol.22, No.8, pp.809–830j, 2000.

[4] W.Grimson, C.Stauffer, R.Romano, L.Lee: "Using A Adaptive Tracking to Classify and Monitor Activities in a Site," CVPR98, pp.22-31, 1998.

[5] Q.Zhou, J.K.Aggarwal: "Tracking and Classifying Moving Objects from Video, Proceeding 2nd IEEE Int. Workshop on PETS, 2001.

[6] R.T.Collins, A.J.Lipton, T.Kanade: "A System for Video Surveillance and Monitoring," VSAM final report, Technical CMU-RI-TR-00-12, 2000.