Motion Detection with Networked Cellular Vision System for Preventing Crime and Security

Hiroyuki Kawai[1], Hisato Kobayashi[2]

[1] Department of Robotics, Kanazawa Institute of Technology, Ishikawa 9218501, JAPAN
hiroyuki@neptune.kanazawa-it.ac.jp
[2] Information Technology Research Center, Hosei University, Tokyo 1028160, JAPAN
h@k.hosei.ac.jp

Keywords: Motion Detection, Networked Cellular Vision System, Security System.

Abstract

In this paper, we propose a new concept for information processing of networked vision sensors as security systems. The networked sensor technology has a potential capability to solve some of our most important scientific and societal problems. But, difficulties of processing are always big problems in case of such huge amount of information acquired by the distributed vision systems. The proposed concept gets a hint from information processing of human hearing organs and compound eyes of insects. By a basic experiment, we confirmed that the proposed concept can be utilized to detect human behavior.

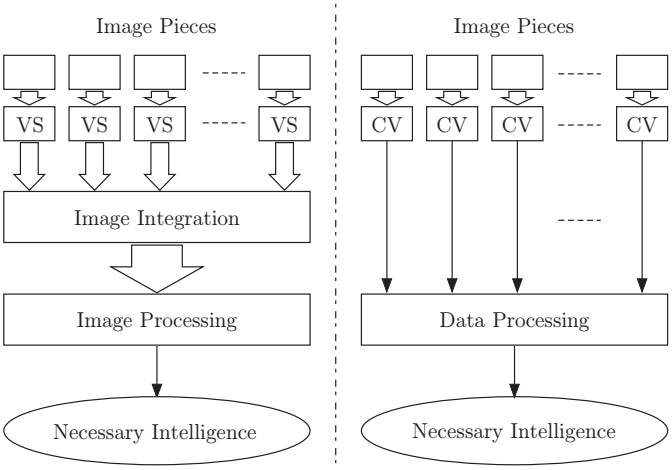


Figure 1: Information flow. Left: With conventional image processing. Right: With information transformation. VC and CV mean a vision system and a cellular vision, respectively.

1 Introduction

Since public spaces are open to everybody, they are also open to dangers lurking in them. In order to prevent criminal matter, it is important to keep monitoring such public spaces: e.g. airport lobbies, railroad stations, parks and so on. However, it is not easy that the guards always check the behaviors of numerous people in such huge public spaces. For these problems, the networked sensor technology has a potential capability to deal this problem. The networked sensors can acquire huge amount of information, especially in case of vision systems. But it has another aspect; if we build large-scale networked sensing system, we face to the serious problems, i.e., how we can handle such huge amount of information and how we can retrieve our necessary intelligence. Even if distributed data processing may alleviate the computational tasks and network traffics; it might be very difficult for the central processor to rebuild and analyze the information of the whole space from the information gotten by decentralized processing [1].

On the other hand, living things processes information very efficiently. Human being recognizes voice or sound by an adroit way. In terrestrial vertebrates, sound waves in the air enter the outer ear, strike the tympanic membrane [2]. The sound waves are converted to fluid waves in the cochlea by a series of mechanical couplings in the middle ear. The fluid waves cause vibration of the basilar membrane, on which sit sensory hair cells in the Corti's organ [4]. Our brain can recognize the sound or the voice by which hair cells are oscillating and how big their magnitude. Namely, our brain retrieves the necessary information from the sound waves by monitoring the dynamical motions of hair cells. In other words, each hair cell compresses the information in the allotted frequency band ideally. The information format is changed from sound to dynamical motion.

There are two aspects.

1. Supervisory Monitoring: The central processor does not treat local data directly; it retrieves necessary intelligence from Meta data acquired by local agents.

2. Changing Information Category: The central processor does not treat image data directly; it retrieves necessary intelligence from different kind of physical value, i.e., motions of vision cells. Namely, the original data is transformed into a different type of physical value.

The first aspect may be rather trivial. It is similar to the concept of distributed processing or decentralized processing; but there are still many difficulties in such processing methods.

The final intelligence can be retrieved only from the merged data. By merging the local data, new synergetic information is born, which is never observed in each local data. Thus, decentralized processing has a limitation, namely the central processor has to play main role. In Fig. 1, the left block-diagram shows the information flow of the conventional image processing and right one shows a new scheme with the information transformation. Moreover, there also exists the similar supervisory monitoring in the natural world. For example, some insects can recognize objects by the compound eyes, although they do not have excellent brains sufficiently [5]. On the other hand, the second aspect is quite new; as we stated above, thanking to such transformation, human being can recognize voices correctly. The process speed of human brain is not so fast, it has so called "ten-step limitation;" it can execute only ten lines source code per second. Even such slow processing mechanism, it can retrieve necessary intelligence from sound waves at every instantaneous moment. Though such concept has potential possibility for information processing, there may not be any engineering applications.

In this paper, by using this concept, we show an example of cellular vision system which can recognize movements of a crowd of people. In a basic experiment, we focus on the human motion detection by the supervisory monitoring.

2 Cellular Vision System

2.1 Basic structure

In this section, we show a case study to facilitate the understanding of our proposed concept. In our case study, we use conventional CCD cameras with pan, tilt and zoom functions. We configure the networked sensing system by connecting large number of these uni-modular CCD devices. Firstly, we show the behavior of the camera model based on the CCD camera. The perspective projection of a target point onto the image plane, $f := [f_x \ f_y]^T \in \mathcal{R}^2$, is given by the following equation.

$$f = \frac{\lambda}{z_c} \begin{bmatrix} x_c \\ y_c \end{bmatrix} \tag{1}$$

where x_c, y_c and z_c represents the target position in x, y and z coordinates of the camera frame [3]. λ is the focal length of the camera and selected as $\lambda = 480$. Let us assume that each camera is hanging from the ceiling of 3[m] height and watching down vertically. Since the resolution of each camera is 320×240 pixels, the camera can watch the area of 2[m] $\times$ 1.5[m] on the floor. We let this area be the responsible monitoring area of the camera. The camera has two modes: normal mode and 3X-tele-mode (three times closer). Even in 3X-tele-mode, we like the camera covers the same responsible monitoring area, 2[m] $\times$ 1.5[m]. This fact makes another assumption that each camera can change its direction within the following ranges.

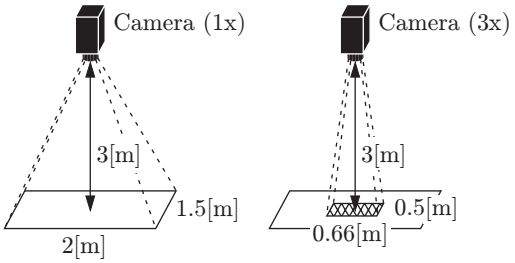


Figure 2: Field of view of the camera in the 3D workspace.

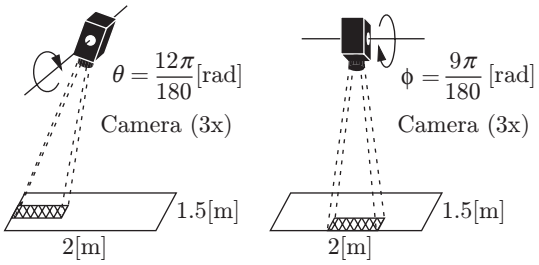


Figure 3: Camera motion in tele-mode.

- Pan: $-12\pi/180 \leq \theta \leq 12\pi/180$ [rad]
- Tilt: $-9\pi/180 \leq \phi \leq 9\pi/180$ [rad]

Fig. 2 and Fig. 3 shows the intuitive illustrations of the above explanation.

We assume that the target point on the image plane is always available without referring the image processing. The following is the camera motion of explaining how to track the target.

1. The initial setting of camera is in normal mode and at the original direction: $(\theta, \phi) = (0, 0)$.
2. If a target gets in the image plane, the camera tries to adjust its direction in order to capture the target at the center of the image plane, or at least within the center area of range $\|f_x\| \leq 38$ and $\|f_y\| \leq 50$.
3. If the target can be captured in the area of $\|f_x\| \leq 38$ and $\|f_y\| \leq 50$, then the camera is switched into 3X-tele-mode.
4. From now on, we call that the camera is in tracking mode. The camera tracks the target with the simple image based feedback control law of $u = -K(f - f_d)$. Where f is the target point, f_d is its desired location, usually at the origin and K is a gain matrix.
5. If the camera loses sight of the target, it is switched back to the normal mode.

6. If the target is still in its sight, the camera repeats the motion from 2), if not, it back to the initial setting 1).

If the camera catches many target points, then it selects the nearest one in a sense of Euclidian norm. In the following section, we carry out two simulations of human tracking.

We adopt the following simple dynamics as the basic model of human walking.

$$m\ddot{x} + \mu\dot{x} = F \tag{2}$$

where m, μ and F represent mass of the human, friction and force in the human walking, respectively. x is the position of the human. m and μ are constants which satisfy $57 \leq m \leq 63$ and $4 \leq \mu \leq 8$, respectively. F takes a random continuous variable during $[-20, 20]$ throughout the simulation. We assume that the human moves independently. Note that the human motion itself is not important in this paper; the crucial issue to be concerned is to detect the outlines of such motions by the proposed method.

2.2 Cellular Vision System in Square Space

In this simulation, we consider a square space as a target monitoring space. We install $17 \times 17 = 289$ cameras at the ceiling of the square as shown in Fig. 4. Since each camera monitors are of 2[m] $\times$ 1.5[m], the total monitoring area is 34[m] $\times$ 25.5[m]. Simulations are carried out by using Matlab and Simulink with VRML(Virtual Reality Modeling Language) Toolbox. Fig. 5 shows a scene of the simulation made by VRML Toolbox with Matlab. Persons can get in and out from everywhere of the square space.

Fig. 6 shows the sample human walking motions generated by the same manner as explained in the former section. There are eight persons walking in the square. The circles mean their starting points and the crosses represent their locations at the edge of the filed or the final time ($t = 60$) In Fig. 6, the mark '*' denotes the camera locations. Each camera tracks the target walking person autonomously when it gets in the tracking mode. Since each camera is driven in two directions: pan and tilt, we can acquire these two driving voltages as the output data of each vision cell. Fig. 7 shows a vector map, where each vector is composed by these two voltages for each camera.

Since this vector map Fig. 7 almost coincides with Fig. 6, we can conclude that the generated sample motion is clearly rebuilt by our cellular vision system. From the output of the cellular vision system, we can easily recognize the situation of the square: how many persons are walking in the area; how fast they are; which direction they are going to. The output of the cellular vision system is just the driving voltages of each camera, thus the amount of data is remarkably small

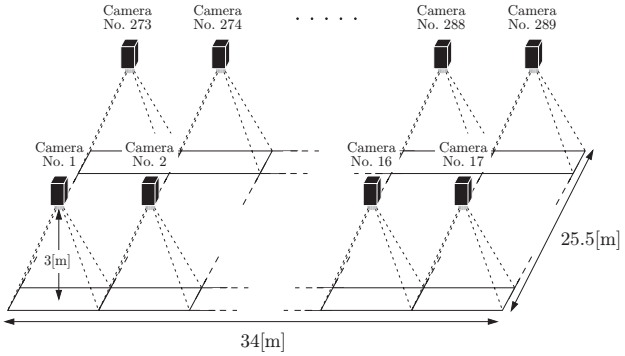


Figure 4: Cellular vision system for square space.

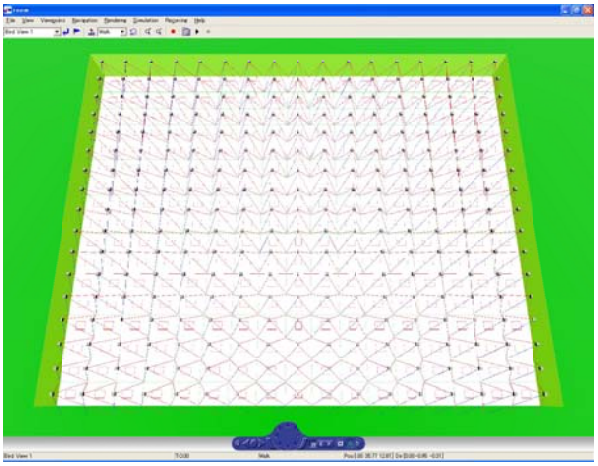


Figure 5: A simulation scene by VRML Toolbox with Matlab.

comparing with conventional vision systems. Moreover, this cellular vision system consists of uni-modular cells, we can easily expand the size of monitoring area just by adding the uni-modular cells.

2.3 Estimation via Parameters of Trajectories

The output vector map of the cellular vision system is very intuitive for human operator. However, if we use this system for a security system, we have to process the data for machine diagnosis, namely to let computers detect irregular situation. The most popular way to describe the trajectories of walking person is polynomial approximation. We assume that the trajectory of the walking is described as the second order polynomial of time t.

$$L_x(t) = a_0 + a_1 t + a_2 t^2 \tag{3}$$
$$L_y(t) = b_0 + b_1 t + b_2 t^2 \tag{4}$$

The speed of the walking is also described as follows.

$$V_x(t) = a_1 + 2a_2 t \tag{5}$$
$$V_y(t) = b_1 + 2b_2 t \tag{6}$$

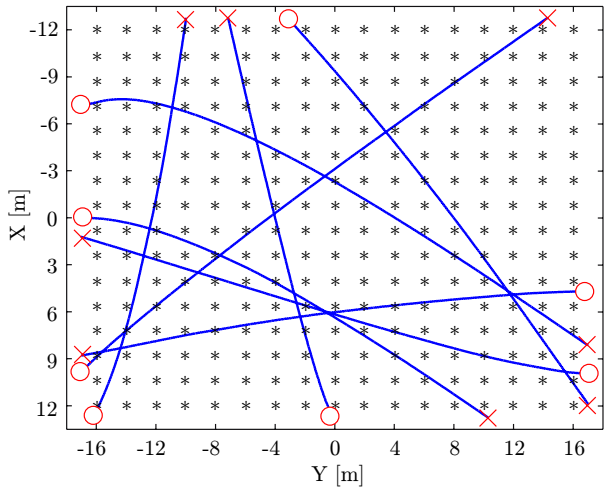


Figure 6: Human motion in square for 60 seconds.

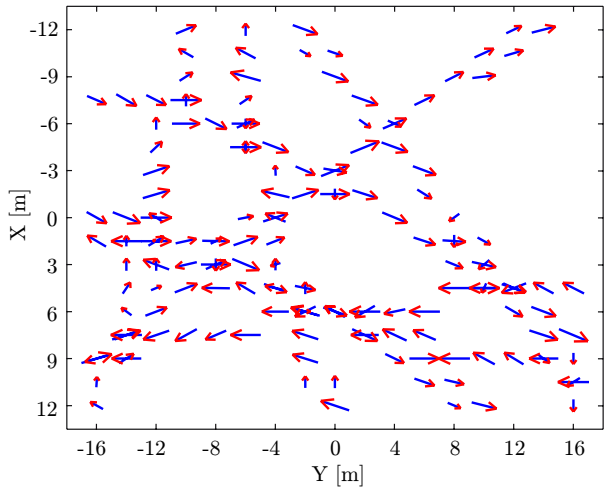


Figure 7: Vector map of driving voltage of each camera

The parameters, a_0 and b_0 only depend on the initial location of the walker.

Based on this assumption, we estimate the trajectories of human normal walking from the vector map as follows.

1. Checking whether there exists an appropriate estimated trajectory for the vector at time k.

2. If there exists the appropriate estimated trajectory, then let the vector be contained in the set which constructs the estimated trajectory. On the other hand, if not, the new trajectory is given for the vector.

3. The estimated trajectories are updated by using the added vectors. Go back to step 1 and set $k = k + 1$.

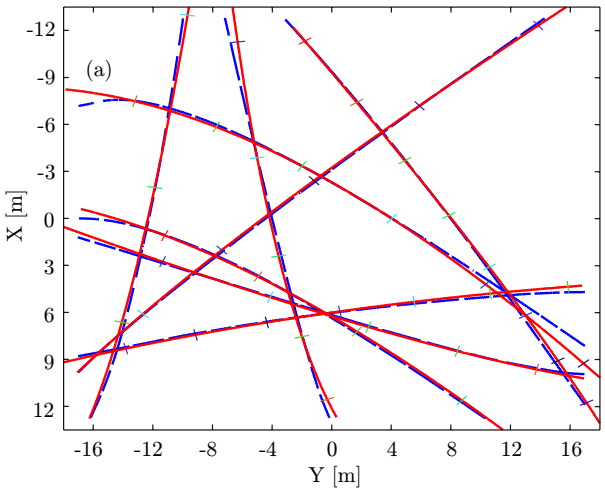


Figure 8: Estimation for human motion in square.

The trajectories are estimated by the least squares fit of the each vector set. The estimated trajectories which can be obtained from the vector map depicted in Fig .7 with the above strategy are shown in Fig. 8. The solid lines and the dashed ones are the estimated trajectories of the human motion and the actual ones, respectively. One of the parameter sets is estimated as follows.

$$L_{xa}(t) = -8.2242 - 0.0962t + 0.0170t^2$$
$$L_{ya}(t) = -23.347 + 0.9584t + 0.0055t^2$$

Though the estimated trajectories do not coincide with the actual ones, they approximate the actual trajectories fairly well. Since human normal walking has some natural properties, the parameters must be in a natural(reasonable parameter) set.

$$(a_1, a_2) \in A_n \tag{7}$$
$$(b_1, b_2) \in B_n \tag{8}$$

These parameter sets are determined from human walking speed and curvature of direction changing.

At any initial time, the system will find several persons in the target space. As time goes, it can search possible trajectories starting from these initial points. The term ”possible” means that the parameters of the trajectory are within the natural parameter sets A_n and B_n. The composed vector of the camera driving voltages implies the tangent vector of this trajectory, thus at any moment, the system can identify the trajectories which the new composed vectors belong to. If a new composed vector (except located on the border of the monitoring area) does not belong to any determined trajectory, then the system recognizes something irregular happens and alert human operators. By this scheme, we can realize the security system for public space, which can alert persons walking in an interrogatory behavior, as well as static data of passenger traffic amount and so on.

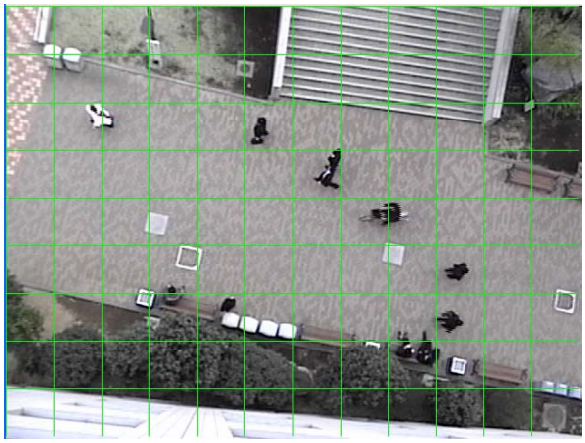


Figure 9: Vision cells with a software approach.

3 Experimental Results with Software Approach

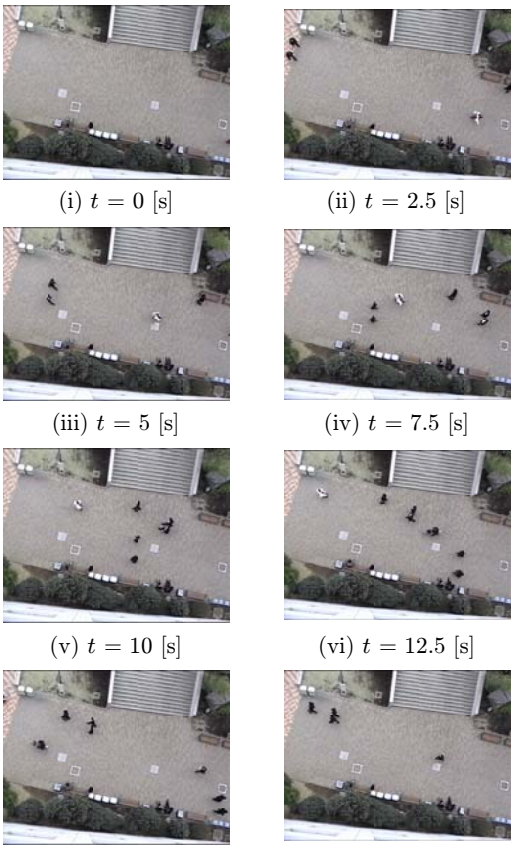
In this section, we show the experiment results focused on the supervisory monitoring only. Monitoring space is about 14[m] $\times$ 18[m] on the campus of Hosei University. Instead of using many camera devices, we adopt only one wide-range camera and divide its image into 108 (9×12) pieces. These pieces are regarded as 108 vision cells for the monitoring space. Each cell monitors about 1.5[m] $\times$ 1.5[m] as shown in Fig. 9.

Fig. 10 shows the actual image sequences of human walking for 17.5[s]. In the interval, there are five people moving from the right hand to the left, two people walking from left to right, one bicycle running from right to left. Fig. 11 and Fig. 12 show the estimated trajectories which can be obtained from the vector maps. While the walking persons of cheek by jowl are occasionally identified as one person (as at (b) in Fig. 12), conventional cases are well identified.

Next, we consider the time sequence of human motion only in the direction of movement as shown in Fig. 13. Clearly, the motion of (c) is fast compared with other entities. Here, we focus on the parameter sets of the direction of movement, i.e., (b_1, b_2). The estimated parameter sets are separated into three groups based on a sign and a value of the parameter b_1 as follows:

$$
\begin{aligned}
B_1 = \{&(-5.5124, 0.0006), (-2.0232, -0.0278) \\
&(-3.7393, -0.0158), (-0.3306, -0.0492), \\
&(-4.2852, -0.0063)\} \\
B_2 = \{&(4.9540, 0.0123), (4.7316, 0.0177)\} \\
B_3 = \{&(105.4913, -1.1352)\}
\end{aligned}
$$

The set B_1 corresponds to the human motion from the right hand to the left, the set B_2 corresponds to from left to right, the set B_3 corresponds to the motion of (c) in Fig. 13, respectively. The parameter set B_3 is not natural compared with other parameter sets B_1 and B_2. Actually, the trajectory of (c) represents the bicycle motion. In this simple experiment, we only exploit the parameter sets of the direction of movement in order to observe abnormal motions.



(i) $t = 0$ [s] (ii) $t = 2.5$ [s]
(iii) $t = 5$ [s] (iv) $t = 7.5$ [s]
(v) $t = 10$ [s] (vi) $t = 12.5$ [s]
(vii) $t = 15$ [s] (viii) $t = 17.5$ [s]

Figure 10: Actual images sequence.

4 Concluding Remarks

Although this paper has shown simple cases of cellular vision systems with experimental results, we can recognize the potential possibility of the idea. The basic concept stated here may be utilized in various fields of information processing. Especially networked sensing systems are getting popular in coming several years; it must be crucial issue to process huge amount of information. The proposed concept may be a key hint to solve these difficulties.

In terms of the cellular vision system with the proposed information processing method, there must be various application fields as follows.

- Security Monitoring for Public Spaces
- Intelligent Transportation Systems (ITS): Surveillance Reckless Driving and Freeway Traffic

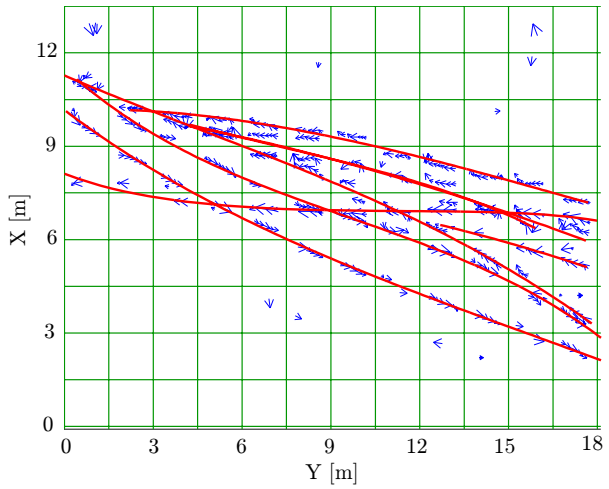


Figure 11: Estimated trajectory with 2nd order polynomial approximation on the monitoring space.

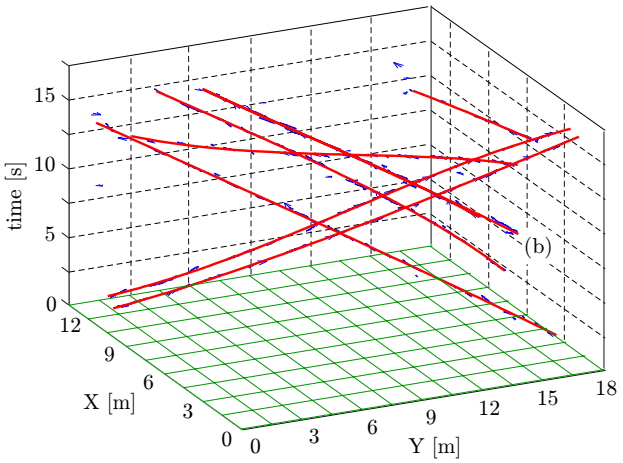


Figure 12: Estimated trajectory with 2nd order polynomial approximation in 3D representation.

- Air Traffic Management by Radar (Commercial and Military use)

In this cellular vision system, each cell has the same vision camera system with the same software, i.e., homogeneous structure; we can produce each cell effectively with affordable cost. Moreover we can increase the number of cells arbitrarily.

Concerning security monitoring for public spaces, we can monitor a huge space for a surveillance. By using the system, we can monitor a crowd and we can easily find a person in an interrogatory manner. This case is same as the case of ITS, we can easily recognize how many cars are running in the specified area and how fast they are running. If a car behaves in abnormal way, we can immediately point out this phenomenon. If we would like to control many objects, airplanes, missiles

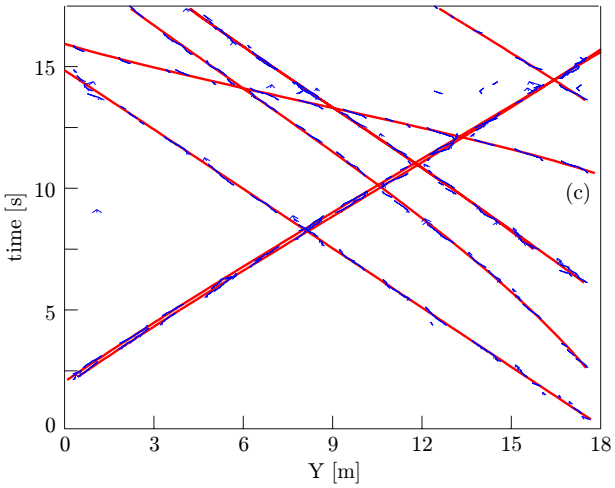


Figure 13: Time sequence of human motion in the main direction of movement.

and so on, flying over tremendous area, monitoring or sensing is absolutely necessary for the controlling. The cellular mini-radar system with the proposed concept may treat this problem.

By the way, the proposed concept may be realized by software rather than the hardware system of uni - modular cell vision system. If we use high resolution CCD camera with fish eye lens and high speed processor, we can emulate hundreds of the cell vision systems. The software realization is cost saving when the target area is covered by a fish eye lens.

Since there are a lot of future works to realize this basic idea in practical systems, we hope some researchers try to use this concept in their developing systems.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. Wireless sensor networks:a survey. *Computer Networks*, 38(4):393–422, 2002.

[2] C. Heneghan, S. M. Khanna, Å. Flock, M. Ulfendahl, L. Brundin, and M. C. Teich. Investigating the nonlinear dynamics of cellular motion in the inner ear using the short-time fourier and continuous wavelet transforms. *IEEE Trans. on Signal Processing*, 42(12):3335–3352, 1994.

[3] S. Hutchinson, G. D. Hager, and P. I. Corke. A tutorial on visual servo control. *IEEE Trans. Robotics and Automation*, 12(5):651–670, 1996.

[4] J. G. Nicholls, A. R. Martin, B. G. Wallace, and P. A. Fuchs. *From Neuron to Brain (4th ed.)*. Sinauer Associates, 2001.

[5] R. Żbikowski. Fly like a fly. *IEEE Spectrum*, 42(11):40–45, 2005.