A Concept of Networked Cellular Vision System for Motion Detection

Hiroyuki Kawai[†] and Hisato Kobayashi[‡]

[†]Department of Robotics, Kanazawa Institute of Technology, Ishikawa 9218501, JAPAN

hiroyuki@neptune.kanazawa-it.ac.jp

[‡]Department of Art & Technology, Hosei University, Tokyo 1028160, JAPAN

Abstract

In this paper, we propose a new concept to process the huge information of networked vision systems. The networked vision systems have potential capability to solve many important problems such as security monitoring of public spaces. But, the information processing of the networked vision system is ”computer-power-consuming task”; it is very difficult to retrieve meaningful information from the data of large size networked vision system. This paper proposes a new simple scheme to process such huge size data, by getting a hint from information processing of human hearing organ and insects' compound eyes. The proposed scheme does not require super computational power, and the necessary computational power is linearly proportion to the number of vision systems. This paper also shows some basic experiments to confirm the validity of the concept.

Index Terms

I. Introduction

The networked vision systems have a potential capability to solve some of our most important scientific and societal problems such as security monitoring for huge public space. Such networked system can acquire huge amount of information, but 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 processors assist the computational task and reduce the network traffic; it might be very difficult for the central processor to rebuild and to 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 [3]. Our brain can recognize the sound or the voice in real time, 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 also exists the similar supervisory monitoring in the natural world. Some insects can recognize flying baits at unbelievable instantaneous moment. They do not have enough computational power in their brain to execute the image processing. Thus, the hint must lay in their compound eyes and the consecutive neural network [4]. Each cell of the compound eyes must send a very simple signal to the neural network; it must not visual image.

We can get two aspects from the natural world.

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. Namely, only by merging the local data, new synergetic information can be born, which is never appeared in each local data. Thus, decentralized processing has a limitation; the central processor has to play main role in any case.

The second aspect is quite new. 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. As we stated above, thanking to such transformation, human being can recognize voices correctly and insects can capture their baits properly.

Based on such observations, this paper proposes a new concept of networked cellular vision

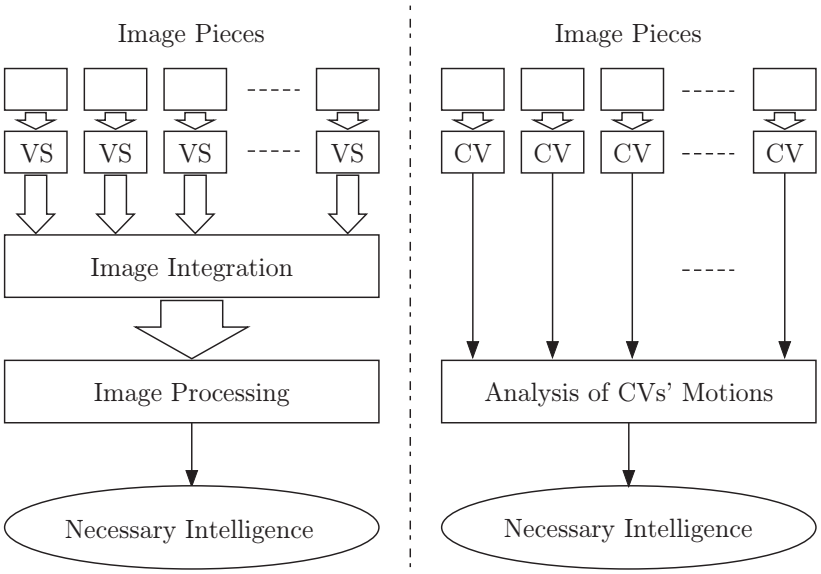

Fig. 1. Information flow. Left: Conventional image processing scheme. Right: Proposed scheme with information transformation. Where VS and CV mean a vision system and a cellular vision, respectively.

processing. In Fig. 1, the left block-diagram shows the information flow of the conventional image processing and right one shows the new scheme with the information transformation; where each cellular vision does not send any image data, it send only the data of its autonomous movement to the central unit. Namely, each cellular vision transforms the image data to a motion data. This transformation plays the main role of simplifying the processing. We show examples of the cellular vision system which can recognize movements of a crowd of people and behavior of microbes.

II. Cellular Vision System

In this section to elucidate the concept of cellular vision system, we explain it by using a simple example. The case we concern here is as follows.

A. Basic Structure of the case study

We use conventional CCD cameras with pan, tilt and zooming 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 [5]. 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.

- 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).

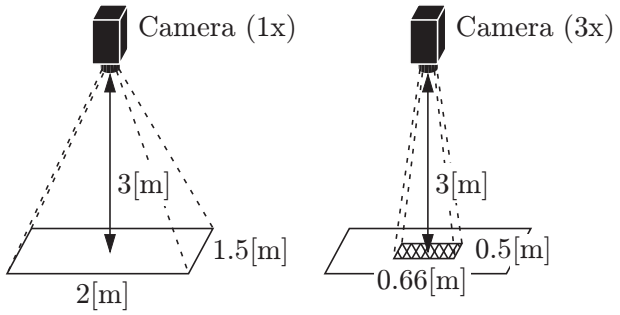

Fig. 2. Field of view of the camera in the 3D workspace.

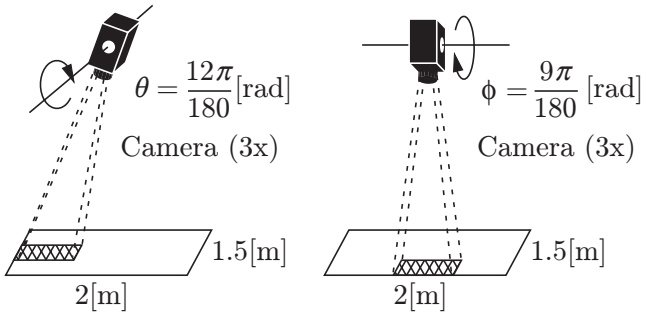

Fig. 3. Camera motion in tele-mode.

If the camera catches many target points, then it selects the nearest one in a sense of Euclidian norm.

B. Simulation of demonstrating the function

In this simulation, we carry out a simulation 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.

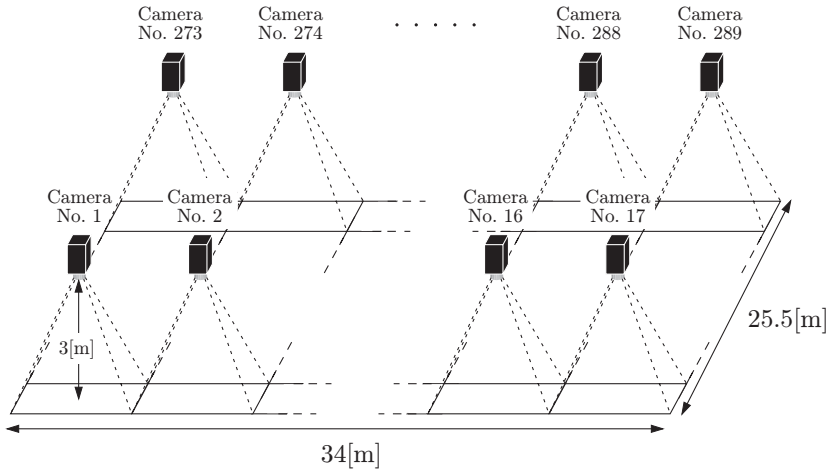

Fig. 4. Cellular vision system for square space.

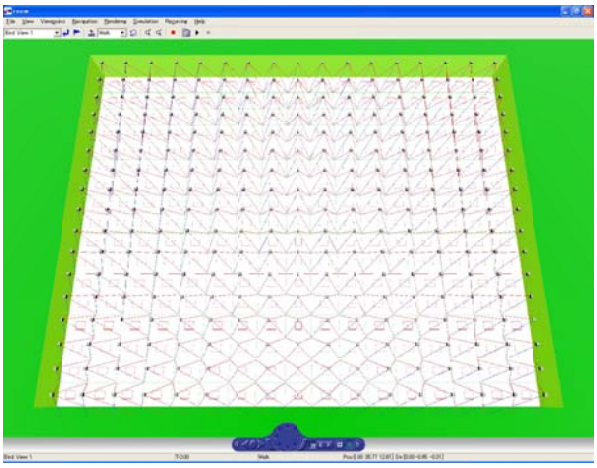

Fig. 5. A simulation scene by VRML Toolbox with Matlab.

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

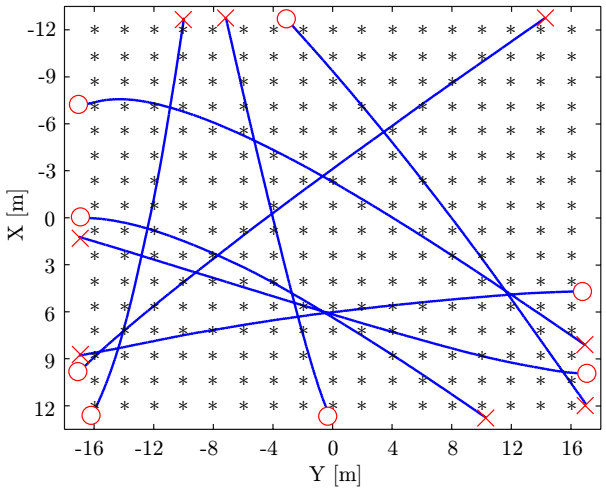

Fig. 6. Human motion in square for 60 seconds.

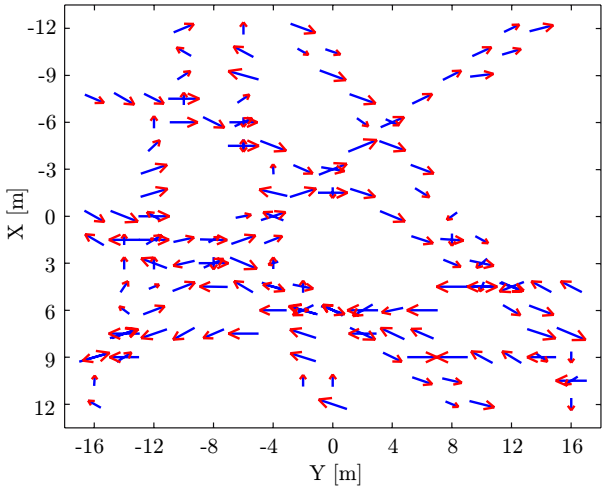

Fig. 7. Vector map of driving voltage of each camera

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 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.

III. Trajectory Finding

As described in the former section, the output vector map of the cellular vision system is very intuitive for human operators. We can easily imagine the continuous trajectory of the moving objects from the discrete vector fields. However, we need more systematic way to determine the trajectory for machine recognition. Each vision cell reports the data of their motion when it finishes the tracking task. The detail is as follows. As described in the former section, when the target objects gets in the range, then the vision cell captures the target and follows it until the target is out of the range. At the timing that the target vanishes, the vision cell reports the following data to the central processing unit.

- Average 2D velocity vector during the tracking.
- Average time of entry time and exit time.

This reporting is ad-hoc event, thus it is event driven and asynchronous reporting. The central processing unit rebuilds the trajectories by the following procedure.

A. Initial Setting

The central processing unit has the following data set, which may be null when the whole system starts.

1) n-trajectories $g_1(t), g_2(t), \cdots, g_n(t)$, which are time functions in X-Y plane and they describe the motions of n objects
2) each trajectory has its basement vector set

$$\mathcal{G}_i = \begin{bmatrix} p_i(t_1) & p_i(t_2) & \cdots & p_i(t_m) \\ v_i(t_1) & v_i(t_2) & \cdots & v_i(t_m) \end{bmatrix},$$

where, $v_i(t)$ is the reported velocity vector at time t and $p_i(t)$ is the corresponding position of the reporting cell. From these vectors, the trajectory is derived by polynomial approximation. Namely, the trajectory was determined as an approximate polynomial whose values and gradients are similar to the data set.

B. Ongoing Processing

Let the k-th vision cell reports the average vector $v_k = \begin{bmatrix} v_{kx} \\ v_{ky} \end{bmatrix}$ and the time t^*. Let the location of the k-th vision cell be $p_k = \begin{bmatrix} p_{kx} \\ p_{ky} \end{bmatrix}$. The central processing unit starts to check the following inequalities to determine whether the new data belong to one of the trajectories.

$$\left| \begin{array}{c} g_{ix}(t^*) - p_{kx} \\ g_{iy}(t^*) - p_{ky} \\ \frac{d}{dt} g_{ix}(t^*) - v_{kx} \\ \frac{d}{dt} g_{iy}(t^*) - v_{ky} \end{array} \right| < \varepsilon, \qquad i = 1, \ldots, n \tag{3}$$

1) YES: If the above inequality is satisfied for j-th trajectory $g_j(t)$, then $g_j(t)$ and its data set $\mathcal{G}_j$ are revised by the reported data p_k and v_k, by defining $p_j(t^*) = p_k$ and $v_j(t^*) = v_k$,

$$\mathcal{G}_j = \begin{bmatrix} p_j(t_1) & p_j(t_2) & \cdots & p_j(t_m) & p_j(t^*) \\ v_j(t_1) & v_j(t_2) & \cdots & v_j(t_m) & v_j(t^*) \end{bmatrix}.$$

Based on the above expanded basement data set, the revised trajectory $g_j(t)$ is derived by polynomial approximation.

2) NO: If the above inequality is never satisfied for any trajectory $g_i(t),\ (i = 1, \ldots, n)$, the central processing unit adds one more new trajectory defined as follows.

$$g_{n+1}(t) = \begin{bmatrix} p_{kx} \\ p_{ky} \end{bmatrix} + (t - t^*) \times \begin{bmatrix} v_{kx} \\ v_{ky} \end{bmatrix} \tag{4}$$

And its data is given by defining $p_{n+1}(t^*) = p_k$ and $v_{n+1}(t^*) = v_k$,

$$\mathcal{G}_{n+1} = \begin{bmatrix} p_{n+1}(t^*) \\ v_{n+1}(t^*) \end{bmatrix}.$$

By doing the above procedure, the central processing unit always has new up to date trajectory set described as mathematical functions.

By using the above stated algorithm, we calculate the trajectories of the moving objects of the example shown in the section II. We restrict the class of continuous function to 2nd order polynomial function. The derived trajectories are shown as the solid lines in Fig. 8. The dashed lines are the actual trajectories of the human motion. Although these lines are not entirely identical, the estimated trajectories can express normal human motions enough.

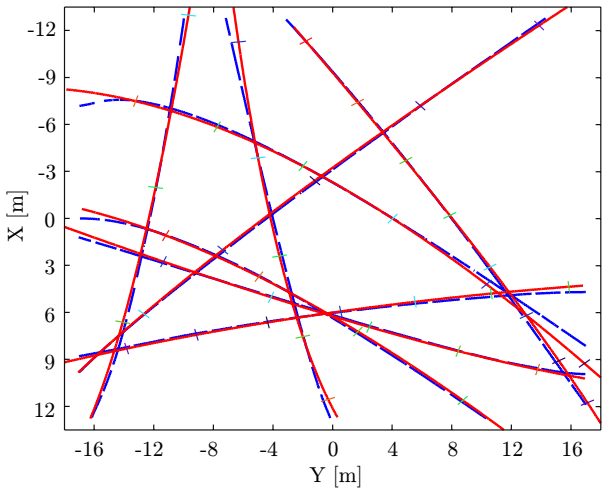

Fig. 8. Estimation for human motion in square.

If we can delete vectors included in the integral sets, from the vector field, the rest is set of noise vector or unidentified vectors. If the density or quantitative amount of such undefined vectors is not negligible, that means something unusual may happen. This criteria may be used for abnormal detection.

IV. Experimental Results

In this section, we show experimental results focused on the supervisory monitoring only. We consider two simple applications based on the proposed concept. Firstly, the monitoring of human motion is shown, similar to the simulation in the previous section. Secondly, the monitoring of microbe behavior is shown. Although the second one is not carried out for a large monitoring space, it gives us the potential possibility of the proposed concept.

A. Monitoring of Human Motion

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 400 (20 $\times$ 20) pieces. These pieces are regarded as 400 vision cells for the monitoring space. Each cell monitors about 0.7[m] $\times$ 0.9[m] as shown in Fig. 9.

Fig. 10 shows the actual image sequences of human walking for 17.5[s]. The original movie can be seen on the Web site [6]. 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

Fig. 9. Vision cells with a software approach.

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 (a) in Fig. 12), conventional cases are well identified.

We also restrict the continuous function class to the 2nd order polynomials. While we could derive 8 trajectories, we categorize these eight trajectories into three classes on the base of the coefficient of the first order terms. These coefficients mean the speed and direction of the moving objects. The following is the set of the coefficients.

$$Group_1 : \{-4.8098, -2.1177, -5.7292, -5.9665, -5.8501\}$$

$$Group_2 : \{5.2344, 4.8684\}$$

$$Group_3 : \{-10.0095\}$$

$Group_1$ corresponds to the human motion of from the right hand to the left; $Group_2$ means the motion from left to right. $Group_3$ is faster than the other group. Actually, the group 3 corresponds to the trajectory of bicycle motion.

B. Monitoring of Microbe Behavior

Like the above stated experience, we can encounter the same situation under the microscope. In order to check the possibility to detect the motion of microbes under the microscope, we tried a very simple experiment. We use a VHX-100 digital microscope with a VH-Z100 lens manufactured by KEYENCE Corporation. In order to catch a microbe by each cell, we

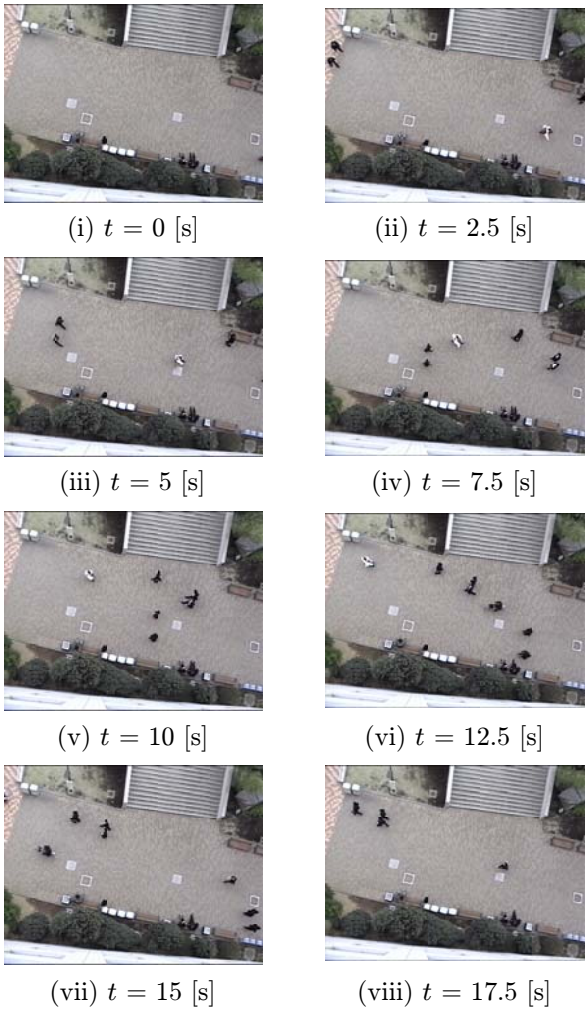

Fig. 10. Actual images sequence.

divide the available region of its image into 1,122 (33×34) pieces as shown in Fig. 13. The original movie from the digital microscope can be seen on the Web site [6]. Comparing with the normal human motion, behavior of microbes is little bit complex. Thus, we restrict the trajectory function class to the third order polynomials.

Fig. 14 shows the estimated trajectories which can be obtained from the vector maps. These two trajectories represent the motions of microbes fairly well. In this experiment, the specimen has only two microbes, thus the necessary computer power and resolution of microscope could be very low. We think that even ordinary setup of the apparatus, we can detect many microbes' motions such as sperm activity of mammals.

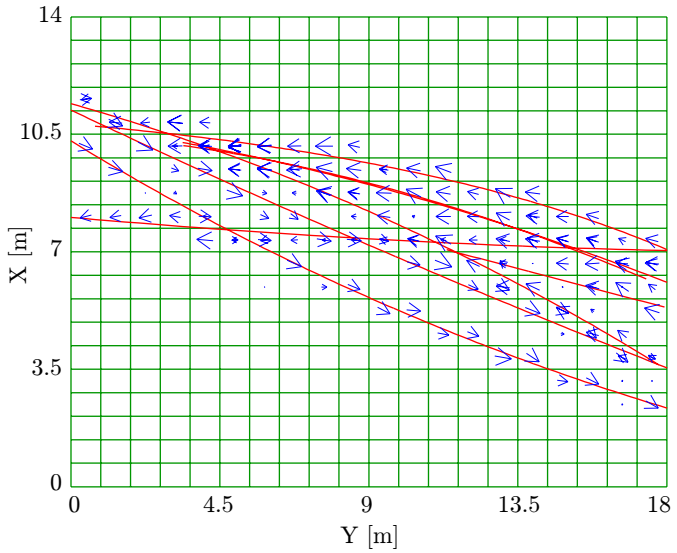

Fig. 11. Estimated trajectory with 2nd order polynomial approximation on the monitoring space.

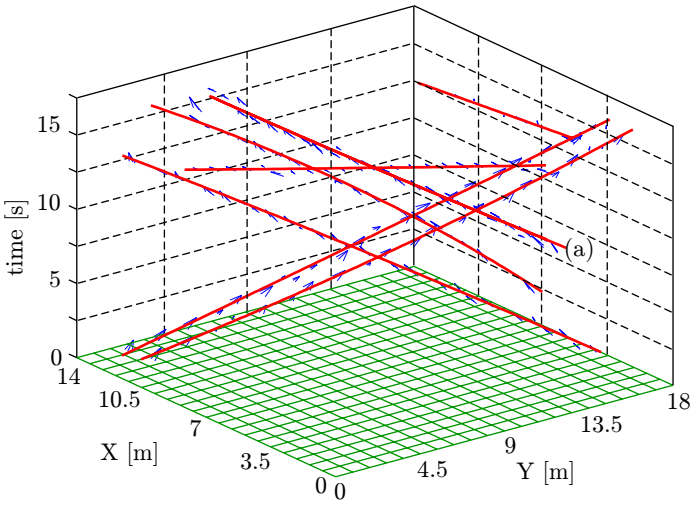

Fig. 12. Estimated trajectory with 2nd order polynomial approximation in 3D representation.

V. 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,

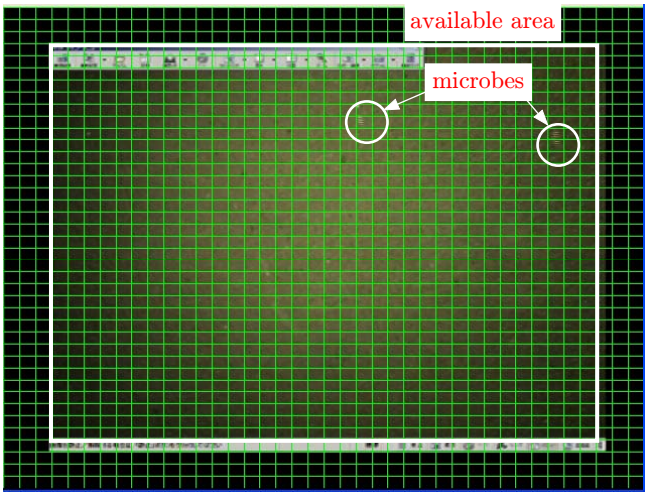

Fig. 13. Two microbes on the image from the digital microscope.

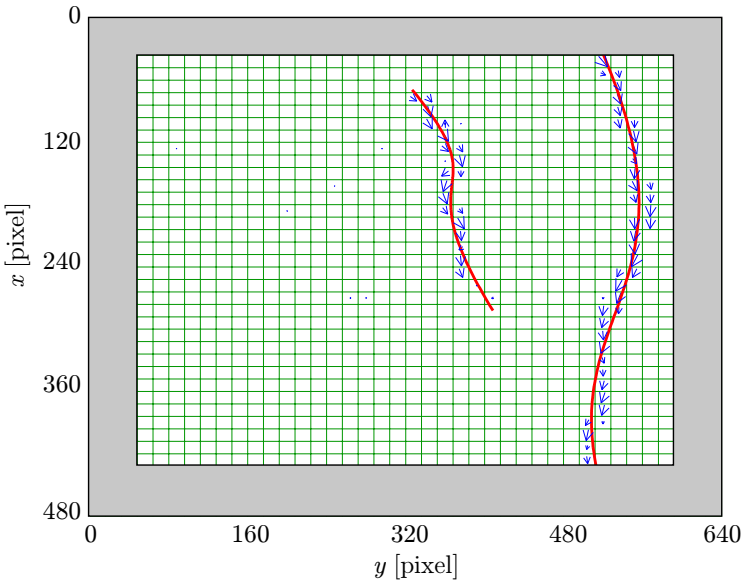

Fig. 14. Estimated trajectories of two microbes with 3rd order polynomial approximation on the monitoring space.

there must be various application fields as follows.

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

ACKNOWLEDGMENTS

The authors would like to acknowledge Prof. Yutaka Tanaka (Department of Art & Technology, Hosei University) and his staff for their extensive help.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, "Wireless Sensor Networks:a Survey," *Computer Networks*, Vol. 38, No. 4, pp. 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*, Vol. 42, No. 12, pp. 3335–3352, 1994.

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

[4] R. Żbikowski, "Fly Like a Fly," *IEEE Spectrum*, Vol. 42, No. 11, pp. 40–45, 2005.

[5] S. Hutchinson, G. D. Hager and P. I. Corke, "A Tutorial on Visual Servo Control," *IEEE Trans. Robotics and Automation*, Vol. 12, No. 5, pp. 651–670, 1996.

[6] http://wwwr.kanazawa-it.ac.jp/kawai/research/NCVS/NCVS.html