A Concept of Networked Cellular Vision System

Hiroyuki Kawai
Department of Robotics
Kanazawa Institute of Technology
Ishikawa 9218501, JAPAN
hiroyuki@neptune.kanazawa-it.ac.jp

Hisato Kobayashi
Information Technology Research Center
Hosei University
Tokyo 1028160, JAPAN
h@k.hosei.ac.jp

***Abstract*— This article proposes a new concept for information processing of networked vision sensors. The networked sensor technology has a potential capability to solve some of our most important scientific and societal problems. But, difficulties of the information processing are inherent in such huge amount of information acquired by the distributed vision systems. The proposed concept gets a hint from information processing of human hearing organs. We get the visual recognition by detecting the physical motions of cell-vision systems that move autonomously in their allotted areas.**

I. Introduction

The networked sensor technology has a potential capability to solve some of our most important scientific and societal problems. 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 effectively. 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 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.

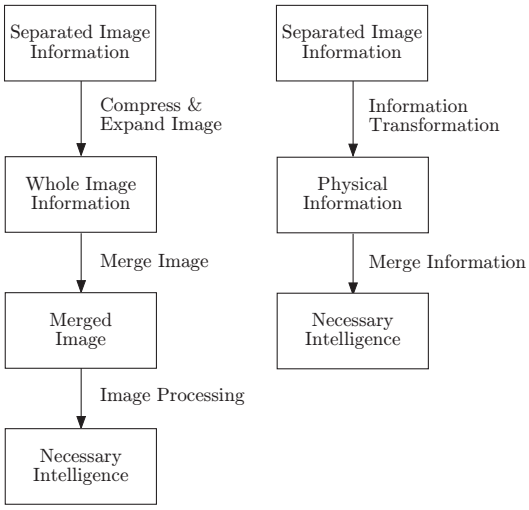

Fig. 1. Information flow. Left: With conventional image processing. Right: With information transformation.

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. 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, the method of which is remarkably easy than the conventional optical flow method.

II. Cellular Vision System

In this section, we show two case studies to facilitate the understanding of our proposed concept. In our case studies, 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 [4]. 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.

In this paper, 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).

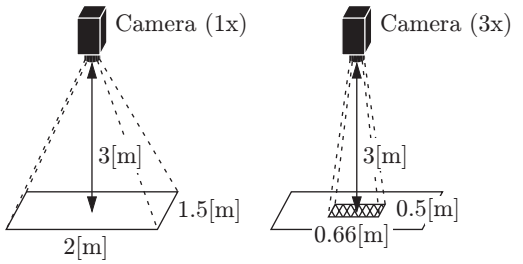

Fig. 2. Field of view of the camera in the 3D workspace.

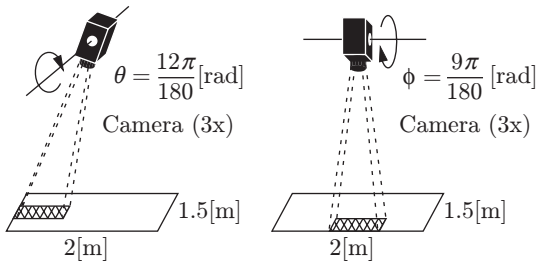

Fig. 3. Camera motion in tele-mode.

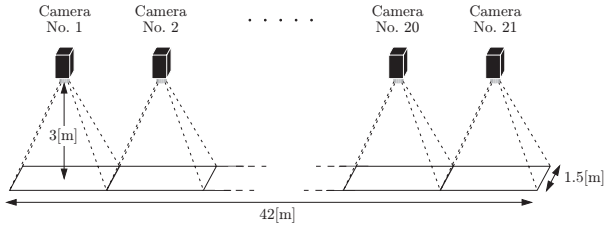

Fig. 4. Cellular vision system in corridor.

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.

A. Cellular Vision System For Corridor

In the first simulation, we consider a case of straight corridor as a monitoring space. We set 21 cameras hanging from the ceiling of the corridors as depicted in Fig. 4.

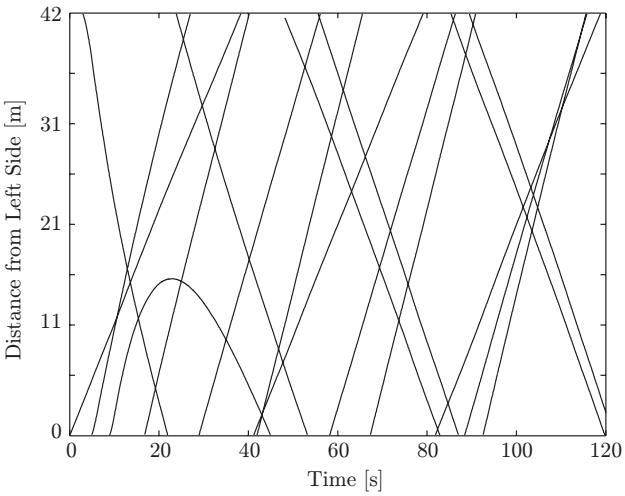

Fig. 5. Human walking in corridor.

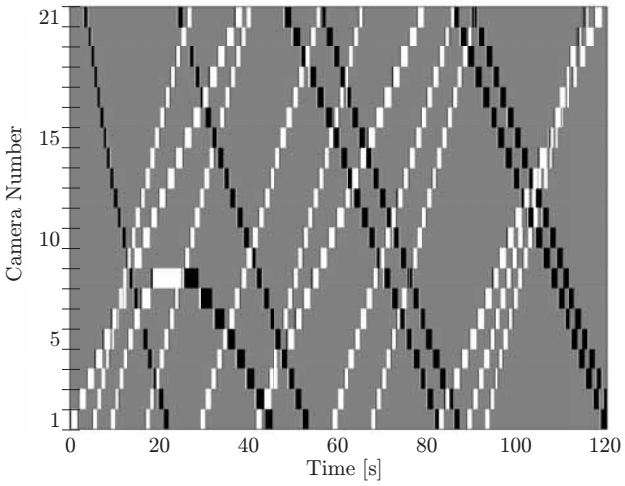

Fig. 6. Sign of the driving voltage on each camera.

Simulations are carried out by using Matlab and Simulink with VRML(Virtual Reality Modeling Language).

We assume that each person goes in and out from the left side or the right side. Since each camera moves only in one direction, the generated movements based on (2) are easily described as in Fig. 5. The horizontal axis is time and the vertical axis is the distance from the left side.

The output from our cellular vision system is described in Fig. 6. It indicates the sign of the driving voltage for each camera. The white mark indicates plus voltage and the black one means minus during the tracking mode. The gray mark is the zero voltage or idling mode (not in tracking mode). The horizontal axis is time and the vertical axis is the camera number.

We can easily conclude that the output of the cellular vision system Fig. 6 coincides with the sample motions of Fig. 5. Namely the cellular vision system can present us the situation of the corridor intuitively. From this output graph, we can easily recognize the facts: how many people are walking; which direction they are walking to; how fast

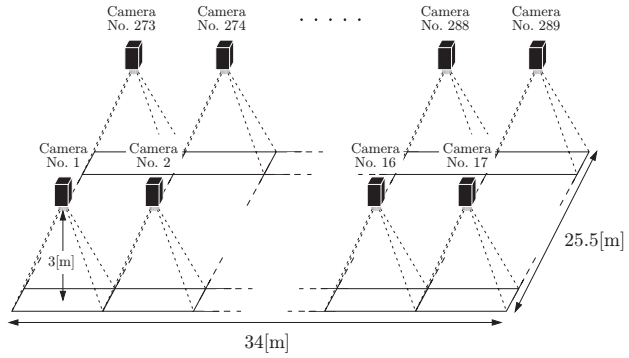

Fig. 7. Cellular vision system for square space.

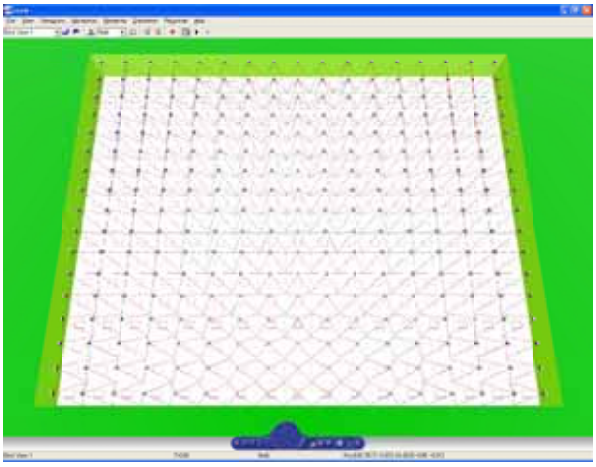

Fig. 8. A simulation scene by VRML.

they are. For example, at the moment of 60 seconds, there are two people moving from the right hand to the left, two people walking from left to right, one just starts walking from left to right.

B. Cellular Vision System in Square

In the second 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. 7. Since each camera monitors are of $2[m] \times 1.5[m]$, the total monitoring area is $34[m] \times 25.5[m]$. Fig. 8 shows a scene of the simulation made by VRML. Persons can get in and out from everywhere of the square space.

Fig. 9 shows the sample human walking motions generated by the same manner as in the corridor case. 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. 9, 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. 10 shows a vector map, where each vector is composed by these two voltages for each camera.

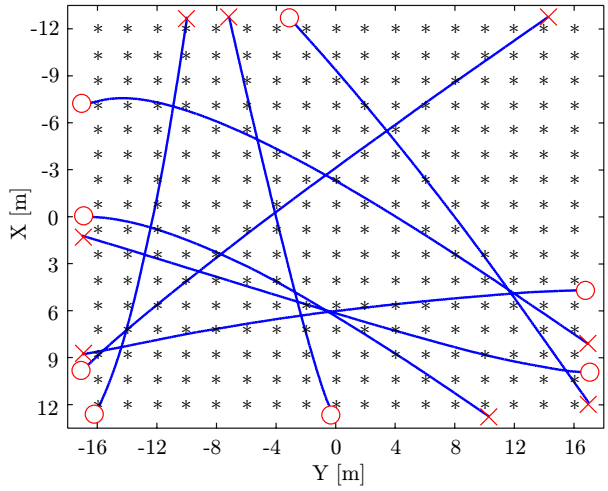

Fig. 9. Human motion in square for 60 seconds.

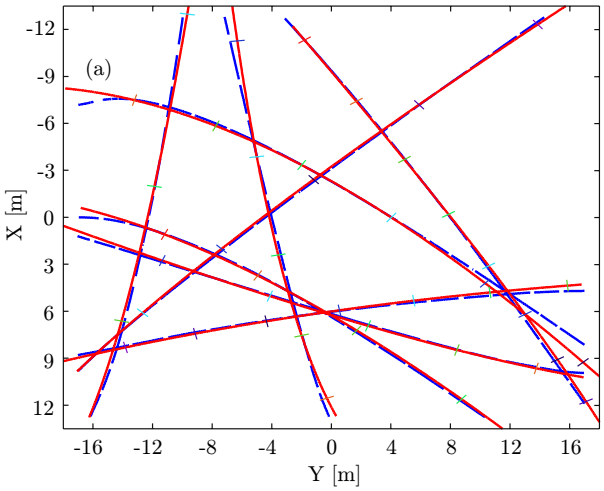

Fig. 11. Estimation for human motion in square.

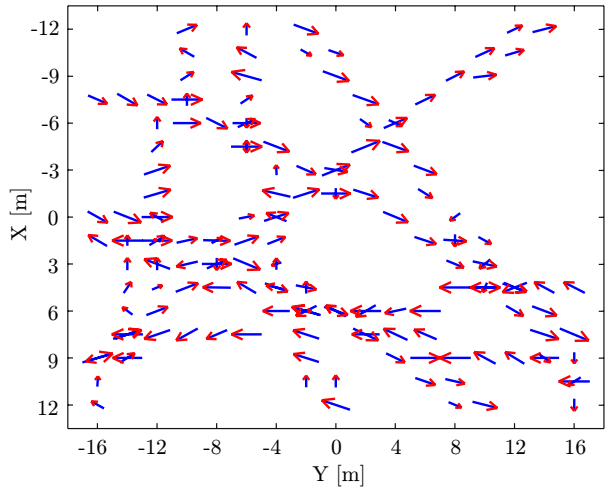

Fig. 10. Vector map of driving voltage of each camera

Since this vector map Fig. 10 almost coincides with Fig. 9, 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.

C. 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 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.

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

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 $k = k + 1$.

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 .10 with the above strategy are shown in Fig. 11. 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.

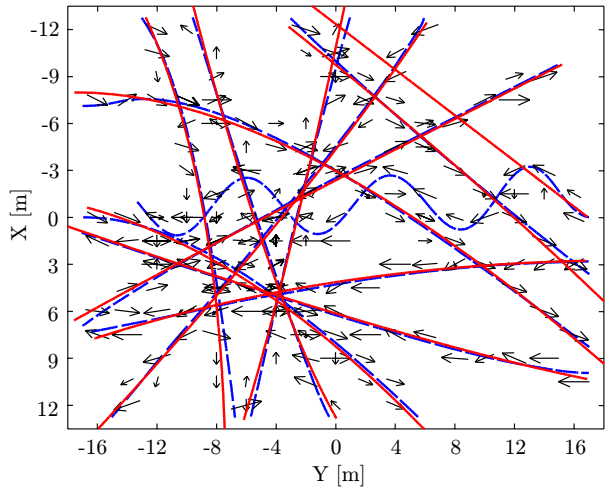

Fig. 12. Estimation with 2nd order polynomial approximation.

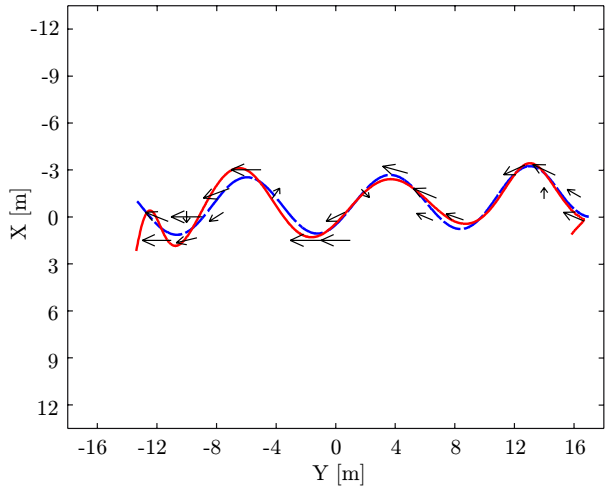

Fig. 13. Estimation with 10th order polynomial approximation.

Next, we give a typical example which includes the meandering motion as an abnormal case. Fig. 12 shows the vector map, the human motions and the estimated ones. Naturally, the abnormal human motion can not be described with the second order polynomial approximation, because the low order approximation is not suitable for complicated motions.

Here, we consider the following function of evaluating the error between the vector and the estimated trajectory.

$$\gamma = \frac{1}{N}\sum_{i=1}^{N} \|(r_{xi} - d_{xi}, r_{yi} - d_{yi})\| \tag{7}$$

where N represents the number of component vectors for the estimated trajectory. (r_x, r_y) and (d_x, d_y) are the positions which are obtained from the vector map and the estimated trajectory at the same time, respectively. If $\gamma \leq 0.55$, then the estimated trajectory is suitable empirically. In order to let the meandering motion satisfy the condition, $\gamma \leq 0.55$, we have to adopt the 10th order polynomial approximation as depicted in Fig. 13. Consequently, we can decide whether there exists the abnormal motion by checking the all values of γ for the estimated trajectories with the 2nd order approximation.

Moreover, we can exploit the parameter set in order to check the abnormal motion. Since human normal walking has some natural properties, the parameters must be in a natural(reasonable parameter) set.

$$(a_1, a_2) \in A_n \tag{8}$$
$$(b_1, b_2) \in B_n \tag{9}$$

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.

III. Hardware Prototype

A. Vision Cell

As we explained, the proposed cellular vision system consists of many uni-modular vision cells; thus this section considers the hardware of the vision cell. The required function of vision cell is only to track the target person when the person gets into its sight. The precise tracking is not required, just a rough tracking is enough for the purpose. The necessary information is driving voltage for the tracking. To realize such simple tracking, visual feedback mechanisms studied widely can be utilized. The feedback law itself is simple and necessary data processing is also simple, thus we can implement this feedback law into a chip-controller. In terms of camera resolution, low resolution is enough for this purpose. However, commercially available CCD chips has too high resolution, usually it is more than 10,000 pixels at least. Thus we can realize the zooming function electronically by using this high resolution in place of optical zooming. Thus, the necessary mechanism is to dive the camera directions: pan and tilt. To realize this mechanism, as actuators, we use four so-called ”artificial muscle” made from high polymer. These four strings of this bind together the corners of camera body and the corresponding positions of the basement, as shown in Fig. 14 and Fig. 15. Since CCD camera can also play the role of angle sensors, we can realize the vision cell only by these actuators.

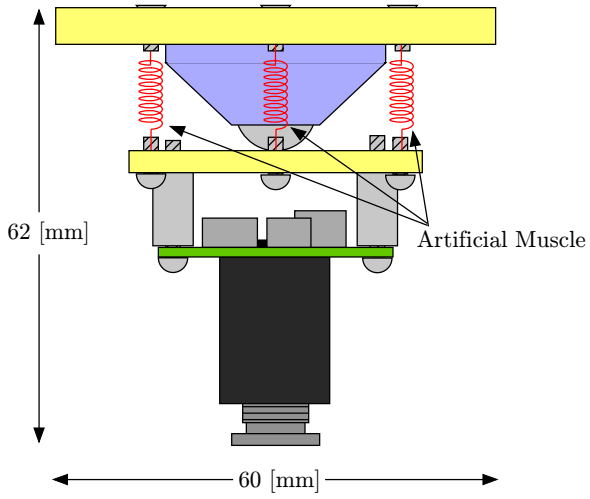

Fig. 14. Model of cell camera.

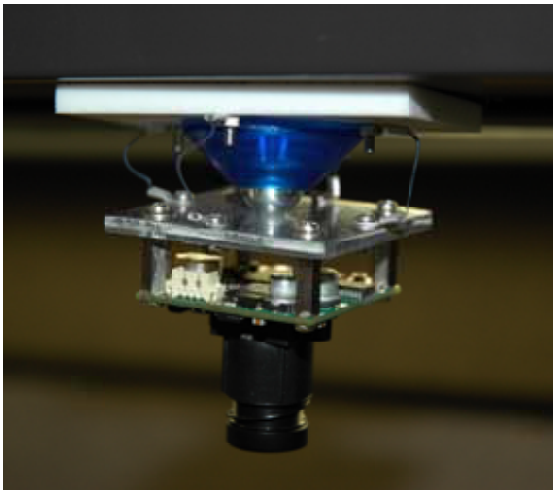

Fig. 15. Prototype of cell camera.

B. Networking

The vision cells are connected by a wire, which transfer their information to the central processor and also provides electric power to the vision cells. As stated before, the information generated by each vision cell is a pair of driving voltages. If the voltage is described with 8-bits signal and sent every second, the amount of information generated by each vision cell is 2 bytes per second. Even adding the ID of vision cell and some protocol bytes to the packet, the size of information packet issued at each vision cell is quite small. Thus, we can connect all vision cells by a single wire serially. Conventional coaxial cables of 10MBPS may handle more than 100,000 vision cells. This consideration concludes that connecting task is realized without break through technologies.

IV. Concluding Remarks

Though this paper has shown only simple cases of cellular vision systems, 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.

- Intelligent Transportation Systems (ITS): Freeway Traffic
- Security Monitoring for Public Spaces
- 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 ITS, we can monitor a huge space for a surveillance. By using the system, 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. This case is same as the case of monitoring public space, we can monitor a crowd and we can easily find a person in an interrogatory manner. 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.

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*, 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] 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.