Proc. 8th International Workshop on Multimedia Information Systems (MIS2002), pp.18-27, Tempe AZ, Oct. 2002.

A System for Visualization and Summarization of Omnidirectional Surveillance Video[†]

Noboru Babaguchi[*] **Yoshihiko Fujimoto**[*]
Kazumasa Yamazawa[+] **Naokazu Yokoya**[+]

* ISIR, Osaka University, Ibaraki, Osaka 567-0047, Japan
{babaguchi, f-fujimt}@am.sanken.osaka-u.ac.jp
[+] Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan
{yamazawa, yokoya}@is.aist-nara.ac.jp

Abstract– In this paper, we propose a media system, named OVISS, which visualizes and summarizes the long-term omnidirectional surveillance video to a viewer. OVISS is capable of visualizing the omnidirectional video given from a vision sensor, called HyperOmni Vision, with 360 degrees view field, by coordinating temporal and spatial representation. The representation media of OVISS is a time line showing the temporal interval when the event happened as well as a spatial map showing the location where the event happened. This system is characterized by 1) event detection through collaborative analysis of video and audio, 2) event based spatio-temporal indexing, 3) at-a-glance visualization that makes it easy to understand the temporal and spatial relations of events, and 4) video summarization for each area or for each event. The basic experimental results show that OVISS is potentially useful for surveillance applications.

1 Introduction

An *omnidirectional vision sensor* (hereafter abbreviated as OVS)[1] is composed of a hyperboloidal mirror and a video camera which are placed face to face in the vertical direction. The OVS is capable of capturing a 360-degree horizontal view field scene at a time. Due to this wide view characteristics, it is expected that the OVS could be promising for many applications of surveillance and monitoring[2]. One of the conventional applications is to detect, in real time, extraordinary events such as suspicious individual's invasion and fire occurrence, and give warning to somebody in a distant place. For this purpose, several systems[3–6] concentrating on real-time event detection have been developed using a single sensor or multiple sensors.

In spite of the advantage of its wide view field, an image given by the OVS has disadvantage of low resolution in using an ordinary video camera. This might make it difficult to know "What is he/she doing?" or "Who is it?" which we really want to know. Let us now consider a new application from a different viewpoint. From the *omnidirectional surveillance video* which the OVS is taking for a long time, we selectively store significant part of the video which is a scene of events, *visualize* its whole contents by linking the temporal and spatial relations,

[†]This work was supported in part by a Grant-in-Aid for scientific research from the Japan Society for the Promotion of Science and also by the Telecommunications Advancement Organization of Japan.

and make a *video summary*. This is viewed as a human oriented application: visualizing the surveillance video in a form which is easy to see. This application can be useful for the following practical cases: investigation of customer's behavior at a super market or traffic at a crossroad or a rotary.

In this paper, we propose a system, called OVISS (Omnidirectional Video Visualization and Summarization System), of visualizing the contents of the surveillance video and making its video summary[7]. Viewing the video as multimedia streams consisting of omnidirectional continuous images and synchronized audio, we make use of techniques of media content analysis and visual interfaces. Specifically, we introduce *event based spatio-temporal indexing* to the surveillance video. It is an indexing about not only when to happen but also where to happen for the event of interest. The representation media of OVISS is a *time line*[7–9], showing the temporal interval when the event happened as well as a *spatial map* showing the location where the event happened. This system is characterized by 1) event detection through collaborative analysis of video and audio, 2) event based spatio-temporal indexing, 3) at-a-glance visualization[7–10] that makes it easy to understand the temporal and spatial relations of events, and 4) video summarization for each area or for each event.

The rest of this paper is organized as follows. Section 2 shows the configuration of OVISS, and Section 3 presents its prototype and discusses its basic performance. Section 4 gives concluding remarks.

2 Configuration of OVISS

This section describes the configuration of OVISS. It consists of several processing modules: sensing, video analysis, image transformation, and visualizing/summarizing modules. We proceed to explain each of the modules.

2.1 Sensing Module

In this application, we should concentrate on the point that a viewer can see images very naturally. The images have to be easily transformed from an omnidirectional image that the OVS takes. Among a lot of OVSs, we select a sensor called HyperOmni Vision[1] satisfying the single viewpoint constraint, which implies that the given omnidirectional image can be projected to an arbitrary image plane. In fact, HyperOmni Vision meets the above point.

Now let S be an omnidirectional video, and F be an image frame contained in S. The size of F is 640×480 and the frame rate is 30 fps. The environment for surveillance is assumed as an indoor environment under the lighting conditions with less changes. Since OVISS is currently at the preliminary stage, we begin with the simple environment. In addition, it is assumed that the environmental model, for example the layout and physical size of the surveillance area, is known. This assumption may be practical for this application. Figure 1 shows the environment we here deal with. There is a single OVS at the center of the area, and there are four doors to enter and exit there at each corner. The area is equally partitioned into four partial areas designated by Area1, ..., Area4. The correspondence between the areas in the spatial map and the regions in F is indicated in Fig.2.

2.2 Video Analysis Module

The video analysis module performs video segmentation in a temporal dimension through event detection, and event based spatio-temporal indexing. Let us first consider the video segmentation. Unlike TV video, the omnidirectional surveillance video has no scene changes because it is given at a fixed location in the environment. Therefore, the meaningful unit in the omnidirectional video can be the temporal interval when an event takes place.

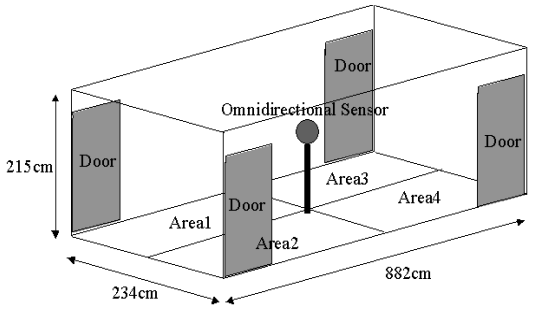

Figure 1: The environment for surveillance.

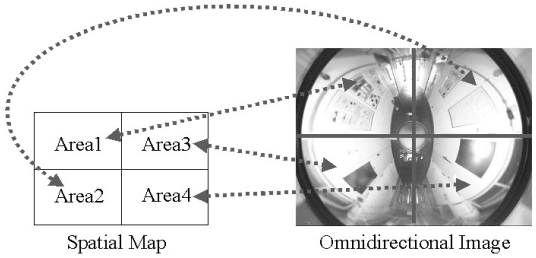

Figure 2: The correspondence between areas and regions in omnidirectional image.

OVISS always recognizes the current state, segmenting the video stream temporally whenever each of the events takes place.

Next we think about what kind of index should be attached to the video. Because the OVS does not have resolution sufficient enough to detect complicated events, we attempt to examine when and where simple events have happened. In this case, as the events of concern, we focus on come-in, come-out, move and stay of a single object assumed as a person. It is noted that the events are defined over temporal intervals. The *temporal index* is based on the media time. On the other hand, the *spatial index* is the name of the area such as Area1, ..., Area4. We call this process event based spatio-temporal indexing.

We now describe how the events should be detected. What should be detected in an environment is the state change induced by the event. Appearance, disappearance and move of an object will certainly cause the state change. To detect these, tracking the object is strongly required. Since we deal with the indoor environment, we exploit a simple image analysis method of background subtraction, which requires a low computational cost.

To perform the background subtraction, an image of a static environment where there is no moving object is captured in advance. Making the difference between the input image frame F and the background image yields a binary image through thresholding. In this image, a region of pixels whose value is equal to 1 may represent the figure of the object. After eliminating noises, OVISS detects the '**Stationary**' state if the number of 1-value pixels is less than a threshold. Otherwise it considers that an object exists and goes into the extraordinary states. In the extraordinary states, OVISS determines the kind, time and area of the occurring event. As mentioned earlier, the events are defined over the temporal intervals. The four kinds of events, i.e. '**In**', '**Move**', '**Stay**' and '**Out**', are as follows.

In: from the time the object appears to the time he/she closes the door.

Move: during the time the object moves in the environment.

Stay: during the time the object stays at the same position.

Out: from the time the object opens the door to the time he/she disappears.

Figure 3 shows the event transition diagram in which the initial state is '**Stationary**.' It can be seen that all the events are related to each other. Each time the image frame is inputted, event states will transit.

To identify the events, we have to investigate the relation between the input image frame F_0 and its neighboring image frames, F_{-1} and F_{+1}. The identification rules are given as follows:

- If F_{-1} is '**Stationary**,' and an object exists in F_0, then F_0 is identified as '**In**.'
- If an object exists in both F_{-1} and F_0, and if the distance between its centroids in F_{-1} and in F_0 exceeds some threshold, then F_0 is identified as '**Move**,' otherwise as '**Stay**.'

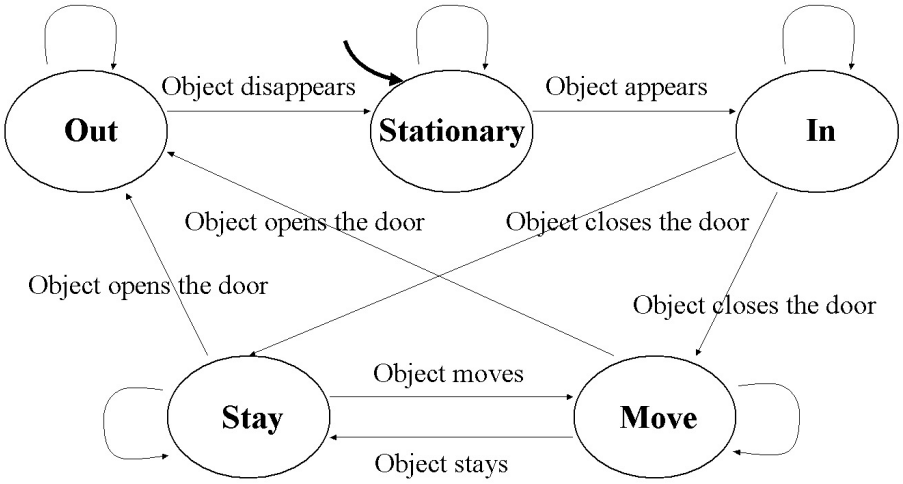

Figure 3: Event transition diagram.

- If an object exists in F_0, and F_{+1} is '**Stationary**,' then F_0 is identified as '**Out**.'

Note that these rules are based on image features. As a result, two kinds of indexes can be obtained. The temporal index is given from the frame number of F_0. The spatial index is the name of the area in which the centroid of the object region is located.

It is, however, difficult to detect whether the door is closed or opened from visual information we can acquire with the OVS. Since the ending frame of an '**In**' event and the beginning frame of an '**Out**' event can not be determined by means of the above rules, we focus on the synchronized audio streams. Namely, we try to identify the image frame by searching the sound when the door is opened or closed. As a metric to express the sound magnitude for an audio frame, we introduce the root mean square (RMS) $v(n)$ of amplitude of auditory signals as

$$v(n) = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} s_n^2(i)}, \tag{1}$$

where N denotes the number of samples included in an audio frame, and $s_n(i)$ denotes the i-th sample value of the n-th frame. The RMS in door open/close is larger than that in the '**Stationary**' state. We determine the image frame with the large RMS as the beginning frame of the '**Out**' event or the ending frame of the '**In**' event.

2.3 Image Transformation Module

When a viewer sees the surveillance video, the omnidirectional image itself is not appropriate for visual inspection. OVISS transforms it into a *planar perspective* image or a *panoramic* image so that the image can be easy for the viewer to see. The former image is equal to the image given with an ordinary camera, and we call simply it the perspective image. The latter means the image of 360 degrees field of view.

Let F, G and R be the omnidirectional image, the perspective image and the panoramic image, respectively. Basically, each pixel value in G or R is computed from the pixel value in F. The transformation scheme is based

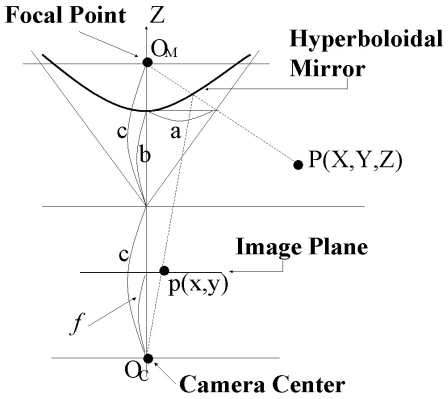

Figure 4: Vertical cross-section of the sensor.

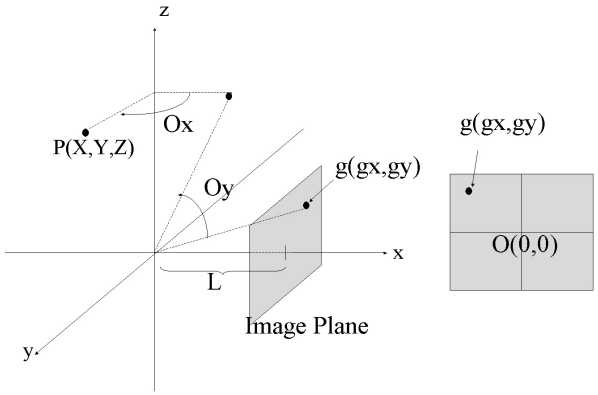

Figure 5: Plane of perspective image.

on the optics of HyperOmni Vision[1, 11]. We first show the transformation of F into G. Let $P(X, Y, Z)$ denote an arbitrary point in the environment. If a point $p(x, y) \in F$ corresponds to $P(X, Y, Z)$, then

$$\tan\theta = \frac{Y}{X} = \frac{y}{x} \tag{2}$$

holds, where θ is the azimuth angle. That is, the azimuth angle of $P(X, Y, Z)$ is equal to that of $p(x, y)$. Considering the vertical cross-section including $P(X, Y, Z)$ and the z axis, as indicated in Fig.4, we obtain the following equations.

$$\begin{aligned} x &= \frac{Xf(b^2 - c^2)}{(b^2 + c^2)(Z - c) - 2bc\sqrt{X^2 + Y^2 + (Z - c)^2}}, \\ y &= \frac{Yf(b^2 - c^2)}{(b^2 + c^2)(Z - c) - 2bc\sqrt{X^2 + Y^2 + (Z - c)^2}}, \end{aligned} \tag{3}$$

where f is the focal length of camera lens, and a, b and c are parameters defining the shape of the hyperboloidal mirror.

Viewing the inner focal point O_M(see Fig.4) as a virtual viewpoint, we map a point $g \in G$ into P in the environment. Now consider generating a perspective image of O_x and O_y degrees in the horizontal and vertical directions, respectively. As shown in Fig.5, we assume an virtual image plane G whose center is $(L, 0, 0)$, where L is the zoom of the generated image. For this image plane, the visual line goes to the point whose latitude and longitude are both 0 degree. Following

$$\begin{aligned} P_x &= (L\cos O_y + g_y \sin O_y)\cos O_x + g_x \sin O_x, \\ P_y &= (L\cos O_y + g_y \cos O_y)\sin O_x - g_x \cos O_x, \\ P_z &= L\sin O_y - g_y \cos O_y, \end{aligned} \tag{4}$$

we can map the point g into P. In these equations, g_x and g_y denote x and y coordinates of g, and P_x, P_y, and P_z denote x, y, and z coordinates of P, respectively. Substituting P_x, P_y, and P_z into X, Y, and Z in Eq.(3) leads to x and y coordinates of p in the omnidirectional image F corresponding to P. For each point in G, the above

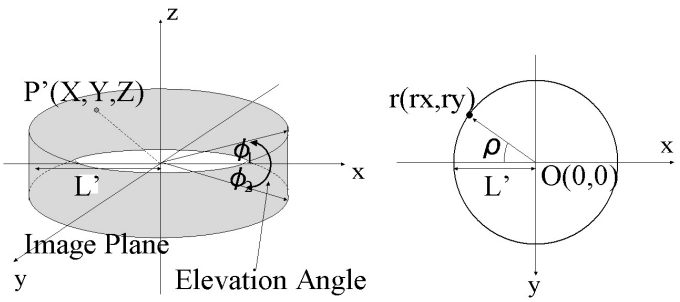

Figure 6: Plane of panoramic image.

transformation process is repeated. Finally the perspective image G where the object is centered is generated. In OVISS, G is of size 320×240.

Next we describe the transformation of F into the panoramic image R. In panoramic images, the zoom is changeable according to the elevation angle, which defines the vertical width with which the image should be generated. We assume the image plane as shown in Fig.6. Here, r_x and r_y represent x and y coordinates of a point $r \in R$. The zoom L' and the angle ρ between the x axis and r are given as

$$\begin{aligned} L' &= \frac{h}{\sin\phi_1 - \sin\phi_2}, \\ \rho &= \frac{-(w/2 - r_x)\cdot 360}{L\pi}, \end{aligned} \tag{5}$$

where h and w are the height and width of R, respectively, and the sum of ϕ_1 and ϕ_2 stands for the elevation angle. Using L', ρ, and r_y, we obtain x, y, and z coordinates of P', denoted by P'_x, P'_y, and P'_z, as

$$\begin{aligned} P'_x &= -L'\cos\rho, \\ P'_y &= -L'\sin\rho, \\ P'_z &= L'\tan\phi_1 - r_y. \end{aligned} \tag{6}$$

Substituting P'_x, P'_y, and P'_z into X, Y, and Z in Eq.(3) yields x and y coordinates of p' in the omnidirectional image F corresponding to P'. As similar to generating the perspective images, the transformation process is repeated for each point in R. In OVISS, R is of size 720(w)×240(h). Figures 7 and 8 show examples of the perspective image and the panoramic image. To avoid the low resolution, a linear compensation technique is introduced for the perspective image.

2.4 Visualizing/Summarizing Module

The visualizing module of OVISS displays the contents, i.e. the events occurring in the surveillance video using the time line and the spatial map. In the time line, the event is marked according to each color. The time line for

Figure 7: Perspective image.

Figure 8: Panoramic image.

each area in the environment is provided. It enables the viewer to understand the transition of event states in the time dimension. On the other hand, the event is also marked on the spatial map. It is noted that the events between the time line and the spatial map are mutually linked. The viewer is able to know the temporal and spatial relation of each event readily.

OVISS is capable of generating the perspective or panoramic image when the viewer selects and clicks the event mark in the time line or the spatial map. In addition, it attaches textual annotation such as <time, area, event name> simultaneously, which is given from the spatio-temporal indexes.

Let us now mention *video summarization*, which is OVISS's main presentation way. It is usually defined as creating shorter video clips or video posters from an original video stream, and is classified into either *spatial expansion* or *temporal compression*. The spatial expansion is to provide image keyframe layouts representing the whole video contents on a computer display like a storyboard[12–14]. It is suitable for at-a-glance presentation. In contrast, the temporal compression is to create a concise video clip by temporally compressing the amount of the video data[15–17]. Its actual examples are movie trails and sports digests.

OVISS summarizes the surveillance video in a temporal compression manner, producing three kinds of video clips, which can be chosen by the viewer, as follows: 1) summary of the specified temporal interval or the whole contents, 2) summary for each area, and 3) summary for each event. In each case, the summary is given as a sequence of either perspective or panoramic images by preserving the original order of event occurrence. Such summaries are intended as a tool to help us understand the global circumstances in a temporal dimension. Gradual scene change operations such as fade-in and fade-out[8] are inserted between the event scenes so that the viewer can easily understand the event change. The viewer can see the video at faster or slower speed, because the playback speed can be controlled.

While OVISS performs surveillance, the background image and the images in extraordinary states are only stored in it. Consequently, a lot of images in ordinary states can be disposed of. Since we assume long-term surveillance in this application, it is not practical to store all the image frames. The above mechanism allows efficient storing of surveillance video.

3 Prototype of OVISS

We have constructed a prototype of OVISS consisting of sensor, video analysis, and interface parts. Video data captured with HyperOmni Vision (HL-W30) and an ordinary video camera was processed at the workstation SGI OCTANE. The video analysis and interface parts were implemented with C and Motif program languages.

Figure 9: Interface of OVISS.

3.1 Interface of OVISS

Figure 9 shows the interface of OVISS. The time line and the spatial map are located in its upper portion. An image sequence and its annotation are displayed in its lower portion, called a *display area*. Each event is colored on the time line and the spatial map. If a viewer clicks the colored area, the perspective or panoramic image sequence corresponding to the event is generated at the display area. In Fig.9, the perspective image is displayed. In addition, OVISS shows the annotation text about the event, synchronizing with the image sequence. The annotation indicates the area name, the event name, and the time interval of the occurring event. The video summary is also shown at the display area. In generating the image sequence, we store an image frame of the transformed image, pasting the frame onto the display area continuously. The time interval between the frames, i.e. the frame rate, is controllable. The time line can be scrolled by a slider, and the temporal interval to visualize and summarize the video can be determined by clicking the slider. OVISS provides the viewer with a dialog that gives the specification in making the video summaries.

In OVISS's demonstration, most of the viewers preferred the perspective image to the panoramic image. We can generate the perspective image in which the object is focused on at the center of the frame. As a result, the sequence of such images looks as if the camera were tracking the object. OVISS, at the current stage, offers only the video summary in the temporal compression type. Combined representation of the time line and image keyframes will produce at-a-glance presentation like the spatial expansion type. We are planning to realize this.

Table 1: Result of event detection and classification.

Events	Numbers	Precision(%)	Recall(%)
In	30	97	93
Move	55	85	84
Stay	24	79	96
Out	30	93	93
Total	139	88	90

3.2 Performance Evaluation of Event Detection

We examined OVISS's performance of event detection and classification under various sensing conditions. The conditions depend on the number of appearing people, the lighting, and the sensor positions. For simplicity, we imposed the following constraint on the experiment: 1) more than two objects do not appear at a time in the environment; 2) the object enters or exits there through the door. We handled five different video streams. The length of each stream was about five minutes.

OVISS terminated the process of event detection and classification in 35 to 45 % of the video length time. Of course, this processing time depends on the number of events that actually took place, and in this case, it was about two minutes. Each average time of generating the perspective images and the panoramic images was 0.36 and 0.44 seconds, respectively. We think efficient processing has been achieved. If we take advantage of the speed-up mechanism[11], real time processing might be possible.

Table 1 indicates the result of event detection and classification for all the five video streams. The performance is evaluated in terms of the recall and precision rates. The recall rate is defined as the rate of the number of correctly detected events to that of actual events. The precision rate is defined as the rate of the number of correctly detected events to that of all the detected events. As shown in the table, almost all the events were detected and classified favorably; the recall and precision rates reached around 90%. The '**Move**' events just after the '**In**' event or before the '**Out**' event were sometimes misclassified as the '**Stay**' events because of less change of object's centroid positions in opening or closing the door. Other reasons for misclassification were the influence of moving objects (people) that incidentally appeared in the surveillance environment and the emergence of spurious regions generated by background subtraction.

4 Conclusions

In this paper, we have proposed OVISS which visualizes and summarizes the long-term omnidirectional surveillance video to a viewer. OVISS realized the visualization from spatial and temporal viewpoints by linking the time line and the spatial map. OVISS can help the viewer understand what kind of events took place in the environment intuitively. In addition, the video summary in the temporal compression manner can be made according to the viewer's requirements. We think that OVISS is a kind of human support system for video surveillance.

Although the prototype of OVISS is at the preliminary stage, the experimental results show that OVISS is promising for surveillance applications. We will extend OVISS's ability by introducing sophisticated methods for image analysis and video processing. Specifically, a multiple-object tracking method robust to the change of lighting conditions should be introduced.

References

[1] K. Yamazawa, Y. Yagi, and M. Yachida: "Omnidirectional Imaging with Hyperboloidal Projection," *Proc. Int. Conf. on Intelligent Robots and Systems*, Vol.2, pp.1029-1034, July 1993.

[2] S.K.Nayar and T.E.Boult: "Omnidirectional VSAM System: PI Report," *Proc. DARPA Image Understanding Workshop*, Vol.I, pp.55-61, 1997.

[3] Y. Onoe, N. Yokoya, K. Yamazawa, and H. Takemura: "Visual Surveillance and Monitoring System Using an Omnidirectional Video Camera," *Proc. ICPR98*, Vol.I, pp.588-592, Sept. 1998.

[4] K.C. Ng, H. Ishiguro, M. Trivedi and T. Sogo: "Monitoring Dynamically Changing Environments by Ubiquitous Vision System," *Proc. Workshop on Visual Surveillance*, pp.67-73, 1999.

[5] R. Miki, N. Yokoya, K. Yamazawa, and H. Takemura: "A Real-time Surveillance and Monitoring System Using Multiple Omnidirectional Video Cameras," *Proc. 4th Asian Conf. on Computer Vision*, pp.528-534, Jan. 2000.

[6] H. Takemura, K. Yamazawa, N. Babaguchi, and N. Yokoya: "Video Surveillance System Using Multiple Omnidirectional Image Sensors," *Proc. Int. Workshop on Pattern Recognition and Image Understanding for Visual Information Media*, pp.81-86, Jan. 2002.

[7] P. Aigrain, H. J. Zhang, and D. Petkovic, "Content–Based Representation and Retrieval of Visual Media: A State–of–the Art Review," *Multimedia Tools and Applications*, 3, pp.179–202, 1996.

[8] A Del Bimbo, *Visual Information Retrieval*, Morgan Kaufmann, 1999.

[9] G. Davenport, T.A. Smith, and N. Pincever, "Cinematic Primitives for Multimedia," *IEEE Computer Graphics & Applications*, pp.67-74, July 1991.

[10] M. Mills, J. Cohen, and Y. Wong, "A Magnifier Tool for Video Data," *Proc. ACM CHI92*, pp.93-98, 1992.

[11] Y. Onoe, K. Yamazawa, H. Takemura, and N. Yokoya: "Telepresence by Real-Time View-Dependent Image Generation from Omnidirectional Video Streams," *Computer Vision and Image Understanding*, Vol.71, No.2, pp.154-165, Aug. 1998.

[12] B-L. Yeo and M. M. Yeung, "Retrieving and Visualizing Video," *Communications of the ACM*, Vol. 40, No. 12, pp.43-52, Dec. 1997.

[13] M. M. Yeung and B-L. Yeo, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content," *IEEE Trans. Circuits and Systems for Video Technology,* Vol. 7, No. 5, pp.771-785, Oct. 1997.

[14] S. Uchihashi, J. Foote, A. Girgensohn and J. Boreczky, "Video Manga: Generating Semantically Meaningful Video Summaries," *Proc. ACM Multimedia,* pp.383-392, 1999.

[15] M. A. Smith and T. Kanade, "Video Skimming and Characterization Through the Combination of Image and Language Understanding Techniques," *Proc. CVPR97*, pp.775-781, 1997.

[16] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video Abstracting," *Communications of the ACM*, Vol. 40, No. 12, pp.55-62, Dec. 1997.

[17] N. Babaguchi, Y. Kawai, and T. Kitahashi: "Generation of Personalized Abstract of Sports Video," *Proc. IEEE ICME2001*, pp.800-803, Aug. 2001.