Proceedings of the 2008 IEEE
International Workshop on Safety, Security and Rescue Robotics
Sendai, Japan, October 2008

Video Data Compression for Robot to Robot Communication

Hideki TAGUCHI
Department of Electrical Eng., Nagaoka University of Technology, Nagaoka-shi, Niigata, 940-2188, JAPAN

Masahiro IWAHASHI
Department of Electrical Eng., Nagaoka University of Technology, Nagaoka-shi, Niigata, 940-2188, JAPAN

Tetsuya KIMURA
Department of System Safety, Nagaoka University of Technology, Nagaoka-shi, Niigata, 940-2188, JAPAN

***Abstract*—** In this report, we propose a new video communication method for the robot vision network in which video signals are shared by "robots" and "humans". A conventional compression algorithm, such as JPEG or MPEG discards component signals insensitive to human eyes, e.g. high frequency band and tiny signals. On the contrary, the proposed method transmits minimum components necessary for recognition by "robots". In case of communications between "robots", data size to be transmitted is reduced by the proposed method. It is also possible to communicate video signals between "humans" by transmitting only the remaining components.

***Keywords*:** *video coding, motion vector, robot vision*

I. INTRODUCTION

The robot vision system has a large variety of applications such as 1) retrieval of environmental information, 2) remote video surveillance and 3) auto localization of a robot. It is becoming popular to apply the stereo-vision, combining with the omni-directional eye, to generate the depth map or to identify location of an obstacle as the environmental information of a robot [1,2]. Recently, the monocular stereo vision systems are also proposed [3].

In the remote video surveillance system, video signals taken by a mobile robot are compressed before it is transmitted. This is because its data size is too large to be communicated via a digital network. A video compression algorithm such as MPEG, JPEG or JPEG-2000 decomposes a video signal into several components. Only some principal components sensitive to human eyes are transmitted to reduce data size for communication [4,5,6].

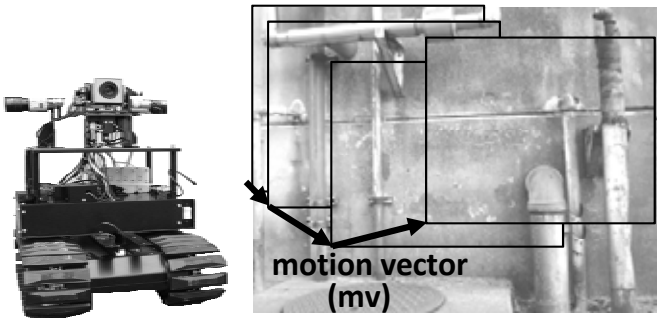
The auto localization system is based on estimation of the ego-motion from video signals. It is called "visual odometry" and effective for the simultaneous localization and mapping (SLAM) technique to build a map of the environment [7,8]. In this case, the optical flow [9] or the motion vector [10,11,12] is estimated from the video signal.

We focus on how to reduce data size for communication in a robot vision network. Several robots and humans are connected by a digital communication network. They are sharing "visual" information. It should be noticed that the principal components sensitive to "human" eyes are not always necessary for "robots".

In this paper, we propose a new layered video coding method which extracts the minimum components necessary for communication between "robots". A video signal taken by a mobile robot in fig. 1 (a) is decomposed into frequency "bands" and "bit planes" by the JPEG-2000 compression algorithm. We examine the minimum components necessary for extracting the motion vector in fig. 1 (b). We also evaluate its data size to be transmitted to the "robots". It is also possible to reconstruct video signals for "humans" by additionally transmitting only the remaining components.

It contributes to smooth communication between "robots" and "humans" under a digital network with limited band width capacity in a rescue scene.

Problem of the conventional method and abstract of the proposed method is summarized in **II**. Details of the proposed method are described in **III**. Effectiveness of the proposed method in respect of data size reduction is confirmed in **IV**. Conclusions are given in **V**.



(a) Mobile robot. (b) Motion estimation.

Fig. 1. Motion vectors (mv) are estimated by a machine from a video signal taken by a mobile robot.

II. Robot Vision Network

A. Existing Method and its Problem

Fig.2 (a) illustrates an existing approach to the robot vision network system. Compressed data of the video signal taken by a camera on a mobile robot in fig. 1 (a) is compressed by the JPEG standard. It is simultaneously transmitted to a remote machine (or a robot) and a human. The machine uses the visual information to estimate the motion vector in fig. 1 (b) to calculate location of the mobile robot. In this case, broad band width (huge data size to be transmitted) is required to the digital communication network. This will cause communication trouble due to lack of data transmission capacity of the digital network.

B. Our Proposal

Fig.2 (b) illustrates the proposed approach. In a mobile robot, the video signal is decomposed into several components. Only the minimum components necessary for estimating the motion vector are transmitted to the remote machine. Therefore, the data to be transmitted for robot-to-robot communication are reduced, comparing to the case where components for human-to-human communication are transmitted in the conventional approach. The video signal is decomposed into several frequency "band" components and "bit plane" components. These decompositions are performed by the discrete wavelet transform (DWT) and the EBCOT in the JPEG-2000.

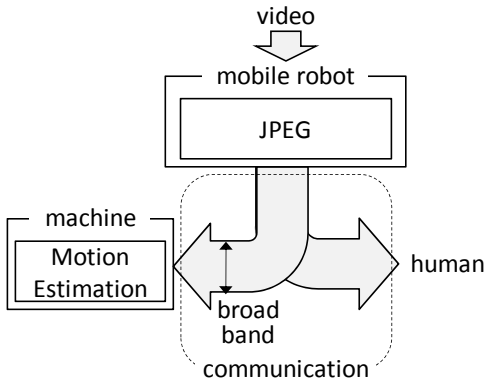
When a human checks the video signal from the mobile robot, only the remaining components are communicated in the same manner of our previously proposed functionally layered video coding methods [13, 14].

III. Proposed Method

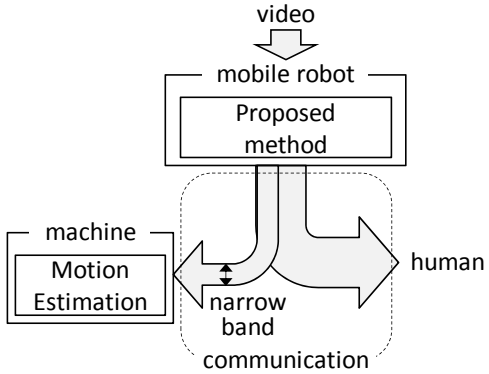
A. Band Decomposition

Fig.3 illustrates procedure of the band decomposition. The input video signal is decomposed into the bands: 1LL, 1LH, 1HL and 1HH in the first stage where L and H demote low and high frequency band signal respectively. In this report, the DWT in JPEG 2000 [6] is used. Each of the band signals are encoded independently (*It is also possible to use the DCT in the JPEG for the band decomposition purpose*). The lowest band signal 1LL is furthermore decomposed into 2LL, 2LH, 2HL and 2HH in the second stage. This is called the octave decomposition. This procedure is repeated to the *n*-th stage.

The minimum bands necessary for extracting the motion vector are categorized into the 1st layer in fig.3. These are examined in **VI**. The remaining bands are in the 2nd layer. These are transmitted to humans as the remaining components as additional data.



(a) Existing method.



(b) Proposed method.

Fig.2. Robot vision network with robots and a human. Data amount to be transmitted to the machine is reduced.

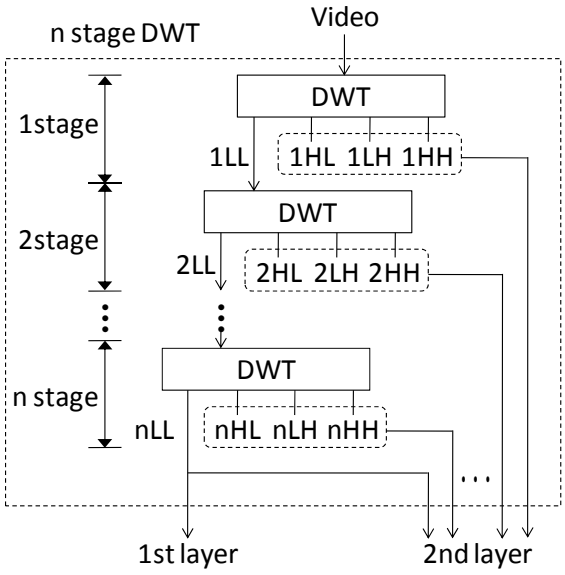


Fig. 3. The band decomposition in the proposed method. The DWT produces band signals LL, HL, LH and HH.

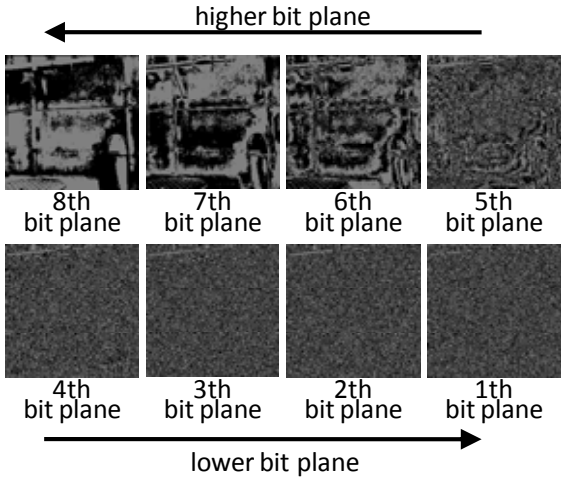


Fig. 4. Examples of the bit plane decomposition. Not necessary to transmit all the planes to a robot.

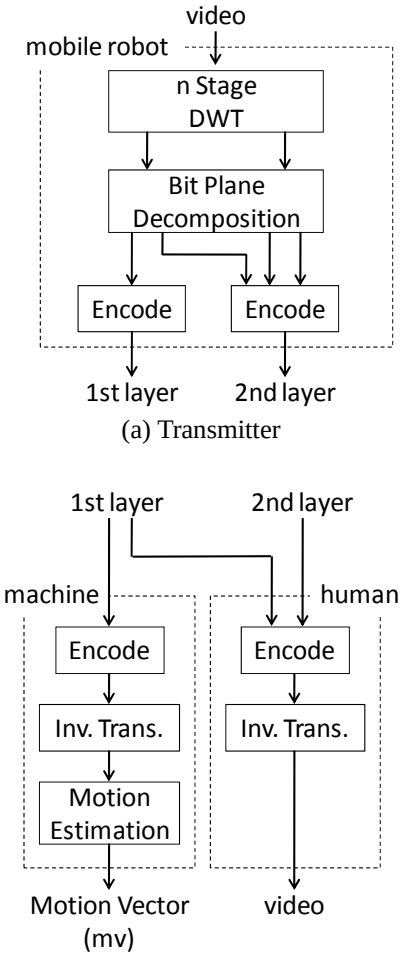


Fig. 5. Signal processing in the proposed method. Extraction and encoding of bands and planes are added to JPEG-2000.

B. Bit Plane Decomposition

Fig.4 illustrates an example of the bit plane decomposition of a video signal in fig. 1 (b). It's value has eight bit per pixel. The 1st plane is the LSB and it indicates that the value is odd or even. The 8th plane is the MSB which indicates that the value is higher than 128 or not. For the robot-to-robot communication, it is expected to be able to estimate the motion vector from a part of these bit planes, so that data amount to be transmitted is reduced.

C. Signal Processing in the Proposed Method

Fig.5. illustrates overall signal processing in the proposed method. In the transmitter, a part of the components necessary for the motion estimation are encoded as the 1st layer bit stream. The machine receives this layer and reconstructs a video signal to calculate the motion vector. The remaining components are categorized in into the 2nd layer. Both of the layers are decoded to reconstruct a video signal for humans in the same manner as the JPEG-2000.

D. Motion Estimation

To extract the motion vector in fig.1 (b), the motion estimation is performed by evaluating one of the criteria defined as below.

(a) CC (cross correlation) [10]

$$E_{CC} = F^{-1}\left[X_1(e^{j\omega_1}, e^{j\omega_2}) \overline{X_2(e^{j\omega_1}, e^{j\omega_2})} \right] \tag{1}$$

(b) SSD (sum of squared difference) [11]

$$E_{SSD} = -\sum_{n_1}\sum_{n_2} \{x_1(m_1, m_2) x_2(m_1 - n_1, m_1 - n_2)\}^2 \tag{2}$$

(c) POC (phase only correlation) [12]

$$E_{POC} = F^{-1}\left[\frac{X_1(e^{j\omega_1}, e^{j\omega_2}) \overline{X_2(e^{j\omega_1}, e^{j\omega_2})}}{\left| X_1(e^{j\omega_1}, e^{j\omega_2}) \overline{X_2(e^{j\omega_1}, e^{j\omega_2})} \right|} \right] \tag{3}$$

In the equations above, the operation $F[\ ]$ and $F^{-1}[\ ]$ are the forward and backward Fourier transform respectively. Intensity of a pixel at the location (m_1,m_2) in the previous frame and the current frame of a video are denoted by $x_1(m_1,m_2)$ and $x_2(m_1,m_2)$ respectively. Their Fourier transforms are expressed as follows.

$$\begin{cases} X_1(e^{j\omega_1}, e^{j\omega_2}) = F[x_1(m_1, m_2)] \\ X_2(e^{j\omega_1}, e^{j\omega_2}) = F[x_2(m_1, m_2)] \end{cases} \tag{4}$$

The motion vector "mv=(n_1,n_2)" in fig.1 (b) is estimated by

$$mv_i = \arg \max_{mv} E_i , \qquad i \in \{CC, SSD, POC\} \tag{5}$$

as the argument which maximizes one of the criteria {E_{POC}, E_{CC}, E_{SSD}}. The minimum components necessary for this estimation are investigated for each of CC, SSD and POC in next section.

IV. Simulation Results

In our experiment, two small sized images with 256 x 256 pixels at random locations are picked up from the image with 360 x 480 pixels illustrated in fig.1 (b). Motion vector is estimated by one of the CC, SSD and POC for these two images. This procedure is repeated 300 times. Difference of the two images is the ideal *mv*. Error between the ideal *mv* and the estimated *mv* are evaluated.

A. Band Components for Robot (1 stage)

Fig.6 indicates the standard deviation of the motion estimation error under the additive white noise on the video signal with the SNR [dB] indicated in the horizontal axis. Fig.6 (a), (b), (c) are for CC, SSD and POC respectively. One of 1LL, 1LH, 1HL, 1HH or all of them is/are used.

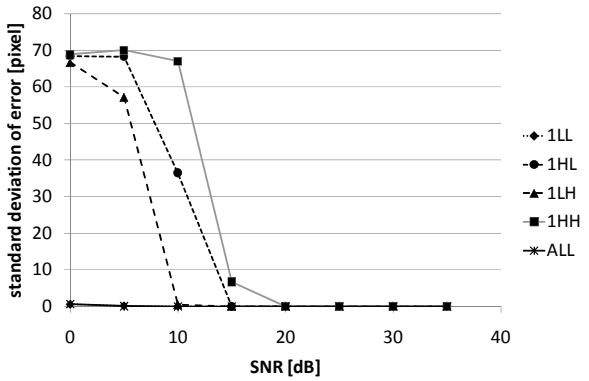
It indicates that the error becomes small under weak noise (high SNR). It was found that there is no difference between 1LL and all bands. It means that not all the bands are necessary but only 1LL is enough to be transmitted to the machine for the motion estimation. It was also found that SSD is the most robust to the noise but POC is the most sensitive. This is because POC uses all the bands equally for the motion estimation. However CC takes importance on variance (energy) and LL has the most in the video signal. It is expected that the data volume is reduced since the number of pixels is reduced to 1/4 by sending only 1LL band instead of all the four bands.

B. Band Components for Robot (n stage)

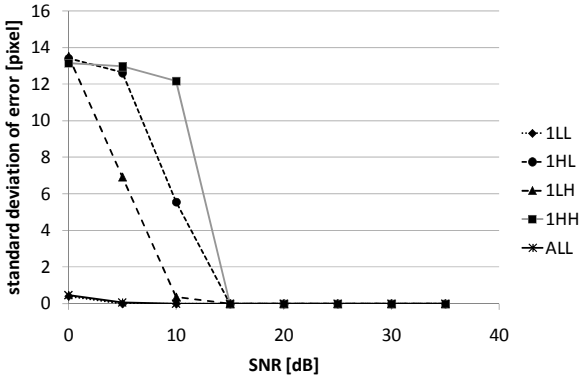
Fig.7 indicates the error in case of *n*LL in the *n*-th stage is used. It indicates that SSD, CC and POC need 2, 1 and 0 stage respectively under the error is zero. Here, 0 stage means all the bands. In the *n*-th stage, the number of pixels of the band component is reduced to 2^{-n} x 2^{-n} so the data volume is.

C. Bit Plane Components for Robot

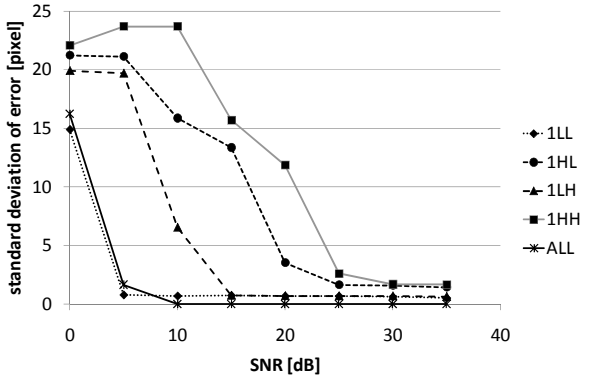
Fig.8 indicates the error in case of a part of the bit planes are used for the estimation. All the bands are used for POC and 1LL is used for CC. In case of SSD, 1LL or 2LL is used. As a result, it was found that only one bit plane (MSB) is enough for precise motion estimation with error is zero for all but SSD with 2LL cases. Data is expected to be reduced to 1/8.



(a) CC (cross correlation)

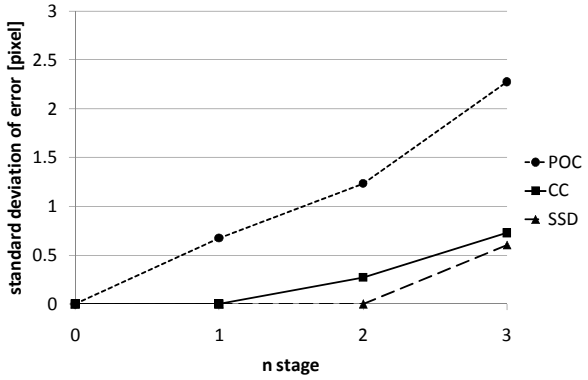


(b) SSD (sum of squared difference)

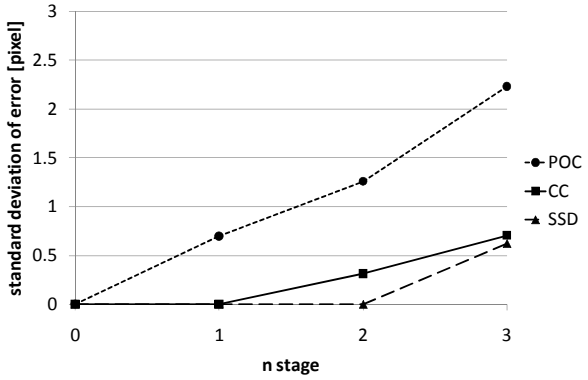


(c) POC (phase only correlation)

Fig.6. Results of the band decomposition. All of {1LL, 1LH, 1HL, 1HH} bands can be replaced by 1LL band.



(a) without noise



(b) with noise

Fig.7. Evaluation of the motion estimation. The SSD is found to be the most robust to the noise.

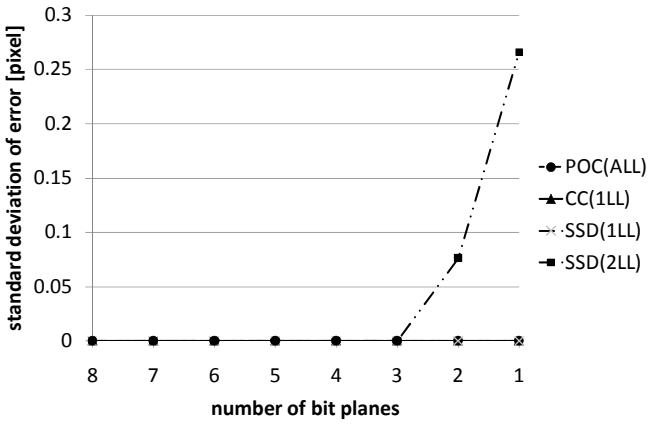


Fig.8. Results of the bit plane decomposition. Only one bit plane is necessary for precise motion estimation.

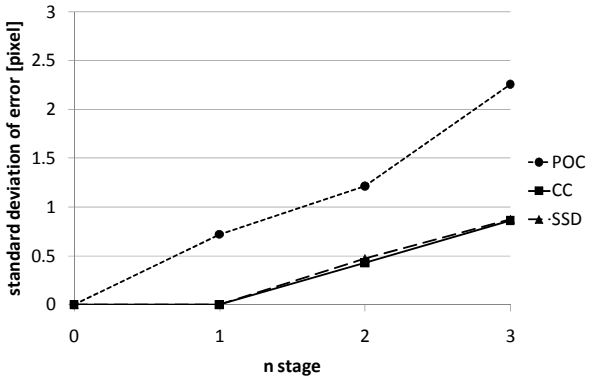


Fig.9. Combination of the band decomposition and the bit plane decomposition.

Table 1 Parameters of the proposed method.

motion estimation	POC	CC	SSD
frequency band	ALL	1 LL	1 LL
number of bit planes	1	1	1
proposed data size[Byte]	4821	1283	1283
existing data size[Byte]	6632	2138	2487
reduction rate [%]	72.7	60.0	51.6

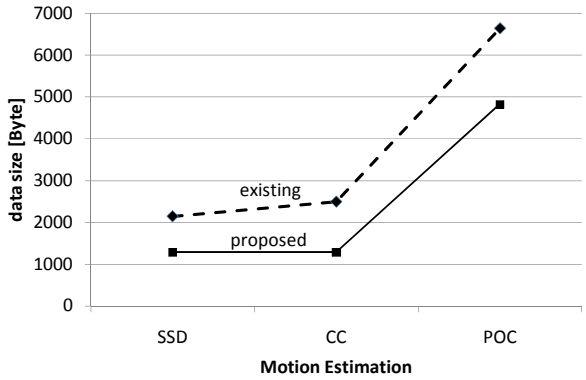
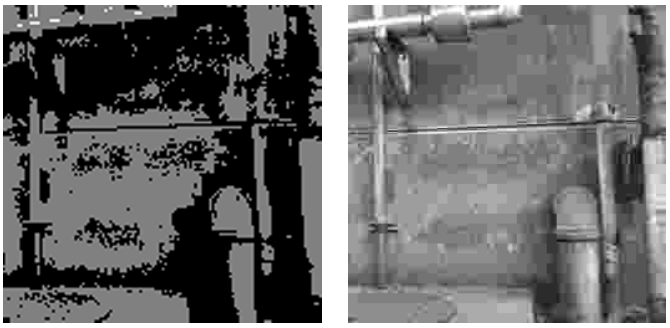
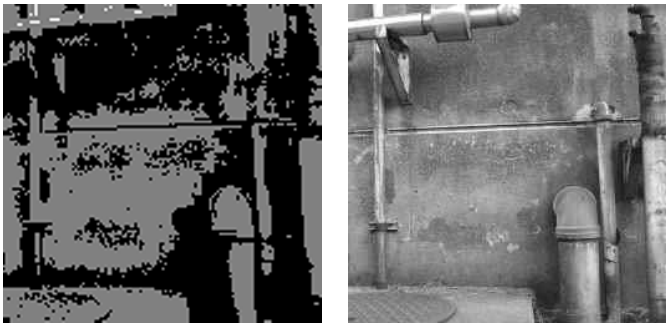


Fig.10. Total data size to be transmitted to the machine.



(a) Proposed (CC,SSD) (b) existing (CC,SSD)



(c) Proposed (POC) (d) existing (POC)

Fig.11. Video signals to be compressed and transmitted.

D. Total Data Size to be Transmitted

Table 1 summarizes parameters determined in this report. Data volume to be transmitted to a robot as the 1st layer is also indicated. It was found that the transmission data is reduced to 51.6 [%], 60.0 [%] and 72.7 [%] for SSD, CC and POC motion estimation methods respectively. These are also illustrated in fig.10. Video signals transmitted to the remote machine for motion estimation are illustrated in fig.11. Since a part of the components are transmitted as the 1st layer, image quality is not preferable for humans as illustrated in fig.11 (a) and (c). However, these are enough for robots and data volume is reduced in case of the robot to robot communications.

When a human observes the video signal, the remaining components are transmitted and then the images illustrated in fig.11 (b) and (d) are reconstructed.

V. Conclusion

In this report, we proposed a new video data compression approach for robot to robot communication via digital network system. We extracted the minimum components necessary for "robot" eyes to estimate motion vectors. It is also possible to communicate video signals for "human" eyes by transmitting additional components. As a result of our experiments, it was confirmed that the transmission data is reduced to 51.6 [%], 60.0 [%] and 72.7 [%] for SSD, CC and POC motion estimation methods respectively.

References

[1] G. N. Desouza, A. C. Kak, "Vision for Mobile Robot Navigation: a Survey", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 24, Issue 2, pp.237- 267, Feb. 2002.

[2] Zhigang Zhu, D. R. Karuppiah, E. M. Riseman, A. R. Hanson, "Keeping Smart, Omnidirectional Eyes on You -- Adaptive Panoramic Stereovision --", IEEE Robotics & Automation Magazine, Vol. 11, Issue 4, pp. 69 - 78, Dec 2004.

[3] T. P. Pachidis, J. N. Lygouras, "Pseudostereo-Vision System: A Monocular Stereo-Vision System as a Sensor for Real-Time Robot Applications", IEEE Trans. Instrumentation and Measurement, Vol. 56, Issue 6, pp.2547 - 2560, Dec. 2007.

[4] JTC1/ SC29, "Information technology -- Coding of audio-visual objects -- Part 10: Advanced Video Coding", ISO/ IEC 14496-10, 2005.

[5] JTC1/ SC29, "Information technology -- Digital compression and coding of continuous-tone still images: Requirements and guidelines", ISO/ IEC 10918-1, 1994.

[6] JTC1/ SC29, "Information technology -- JPEG 2000 image coding system: Core coding system", ISO/ IEC 15444-1, 2004.

[7] D. Nister, O. Naroditsky, J. Bergen, " Visual Odometry ", Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 27, pp. I-652 - I-659, July 2004.

[8] R. Munguia, A. Grau, "Monocular SLAM for Visual Odometry", IEEE International Symposium on Intelligent Signal Processing (WISP), pp.1 - 6, Oct. 2007

[9] M. T. Coimbra, M. Davies, "Approximating Optical Flow within the MPEG-2 Compressed Domain", IEEE Trans. Circuits and Systems for Video Technology, Vol.15, Issue 1, pp.103 - 107, Jan. 2005.

[10] C. A. Wilson, J. A. Theriot, "A Correlation-Based Approach to Calculate Rotation and Translation of Moving Cells", IEEE Trans. Image Processing, Vol. 15, Issue 7, pp. 1939 - 1951, July 2006.

[11] N. P. Papanikolopoulos, P. K. Khosla, T. Kanade, "Visual tracking of a moving target by a camera mounted on a robot: a combination of control and vision", IEEE Trans. Robotics and Automation, Vol. 9, Issue 1, pp. 14 - 35, Feb. 1993.

[12] K. Ito, H. Nakajima, K. Kobayashi, T. Aoki, T. Higuchi, "A Fingerprint Matching Algorithm Using Phase-Only Correlation", IEICE Trans. Fundamentals, Vol.E87-A No.3 pp.682-691, March 2004.

[13] M. Iwahashi, "Awareness Communication Based on Functionally Layered Coding", Picture Coding Symposium (PCS), WedPM3, pp.65-68, Nov. 2007.

[14] S. Udomsiri, M. Iwahashi, S. Muramatsu, "Functionally Layered Video Coding for Water Level Monitoring", IEICE Trans., Fundamentals, Vol. E91-A, No.4, pp.1006 -1014, April 2008.