FUNCTIONALLY LAYERED CODING FOR ROBOT VISION NETWORK

Masahiro IWAHASHI Hideki TAGUCHI Tetsuya KIMURA

Nagaoka University of Technology

ABSTRACT

This paper proposes a functionally layered video coding for robot vision network. Video signals from mobile robots are regularly transmitted to a remote robot. It uses the video to estimate locations of each of the mobile robots. As the need arises, a person browses the video to check the scenery. In such occasion, the functionally layered video coding regularly transmits the minimum components necessary for the motion estimation by the remote robot. For scenery browsing, only additional components are transmitted on demand. As a result, data volume to be transmitted is reduced in total. The proposed method decomposes the video into frequency bands and bit planes by means of the JPEG 2000. This report investigates the minimum components for the motion estimation in the robot to robot communication. It is confirmed that the proposed method can reduce total data volume of the robot vision network.

Index Terms— layered, coding, vision, robot

1. INTRODUCTION

Recently, various applications of the robot vision such as the remote sensing and the auto-localization of a robot have been developed. The omni-directional eye and the stereo vision are utilized to generate a distance map and to identify location of obstacles [1,2,3].

The auto localization technique extracts movement of a robot (ego-motion) from video taken by a camera mounted on it. It is becoming a key technology to perform the SLAM (simultaneous location and mapping) [2,3]. In this case, it is necessary to estimate the motion vector (motion estimation) of frames in the video [4,5,6].

In the robot vision network, some mobile robots with a camera send their video signals to a high performance computer (a remote robot). The remote robot regularly extracts movement of each of the mobile robots to identify their locations. It also offers video scenery of a mobile robot when a person checks it as the need arises.

In this case, the remote robot needs to receive only the minimum components of the video signal necessary for the motion estimation. However, the existing coding algorithms such as the international standards MPEG, JPEG and JPEG 2000 extract components suitable for human eyes [7].

In this paper, we propose a functionally layered video coding for the robot vision network. We extract the minimum components necessary for a robot to accomplish a task. This is a frame work of the functionally layered video coding previously proposed for the privacy conscious communications and the river monitoring [8,9]. In this paper, the task means the motion estimation (ME).

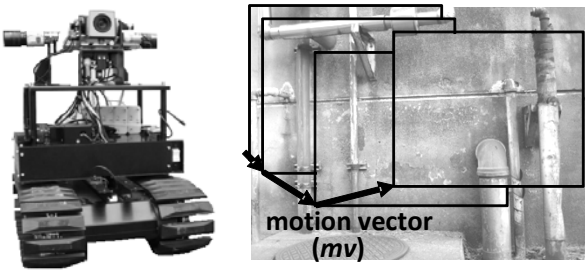
The proposed method decomposes the input video signal into frequency bands F_r by the discrete wavelet transform (DWT), and the bit planes B_p by the bit plane decomposition (BPD). These components are encoded by the EBCOT (Embedded Block Coding with Optimal Truncation) so that the method is implemented by the JPEG 2000 IP core [7].

When all of the components $F_r \times B_p$ are synthesized, the input video is reconstructed without significant loss. In the proposed method, only $f_r \times b_p$ of a subset $f_r \subset F_r$ and $b_p \subset B_p$ are transmitted to the remote robot regularly. As a result, the redundancy in the bit-stream to be sent to the robot is removed and the compression ratio is improved.

2. ROBOT VISION NETWORK

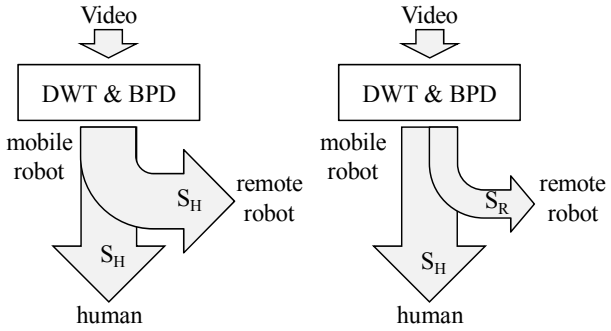
2.1. Problem of Conventional Approach

A video signal taken by a camera fixed to the mobile robot in Fig.1(a) is transmitted to the remote robot. The remote server robot estimates the motion vector *mv* in Fig.1(b) between frames in the video from a mobile robot.



(a) A mobile robot (b) Motion estimation

Fig.1 Motion vector *mv* is estimated by a remote robot for the use of auto-localization of each of mobile robots.



(a) Existing method (b) Proposed method

Fig.2 Video coding is specialized for robot to robot communication. A remote robot receives the minimum components necessary for the motion estimation.

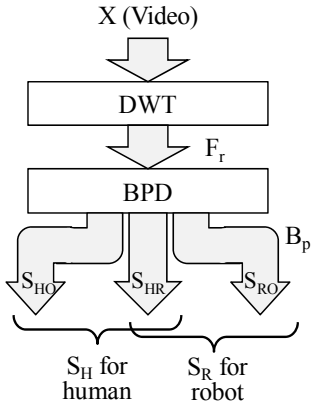
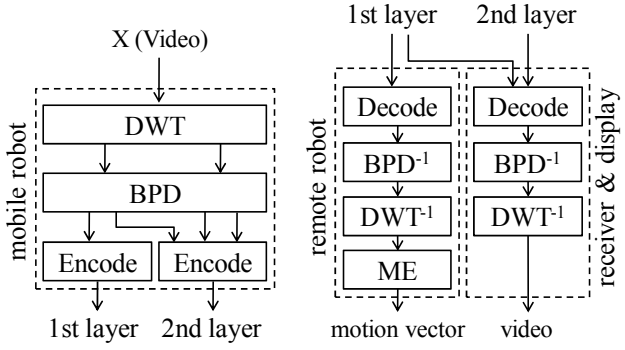


Fig.3 Video signal is decomposed into frequency bands F_r and bit planes B_p. The remote robot receives S_R and a person receives S_H.



(a) Transmitter (b) Receiver

Fig.4 The 1st layer contains S_{HR} for the remote robot. The 2nd layer contains S_{HO} just for a person. ($S_{RO}=\phi$)

In the conventional approach in Fig.2(a), the set S_H (a set of components for *H*uman) composed of a subset of $F_r \times B_p$ is produced by the JPEG 2000 and transmitted to a person for occasional browsing. The problem to be discussed here is that the same set S_H is transmitted to the remote robot and how to remove its redundancy.

2.2. Idea of the Functionally Layered Coding

The proposed method in Fig.2(b) transmits the set S_R (a set of components for *R*obot) necessary for the motion estimation by the remote robot. We expect that the data size of S_R denoted by $DS[S_R]$ becomes less than that of S_H denoted by $DS[S_H]$. Defining the data reduction rate by

$$\eta = \frac{DS[S_R]}{DS[S_H]}, \tag{1}$$

it is expected to be $\eta<1$. Defining the sets in Fig.3 by

$$\begin{cases} S_{HR} = S_H \cap S_R \\ S_{HO} = S_H \setminus S_{HR} \\ S_{RO} = S_R \setminus S_{HR} \end{cases}, \tag{2}$$

the mobile robot sends data volume of $DS[S_{HR}]+DS[S_{RO}]$ to the remote robot regularly, and $DS[S_{HR}]+DS[S_{HO}]$ to a person occasionally. In case of $S_{RO}=\phi$, it transmits the component set S_{HR} regularly in the 1st layer in Fig.4, and it transmits the set S_{HO} occasionally as the 2nd layer when a person checks the video scenery as the need arises.

In this paper, we investigate the minimum components to be included in S_R and confirm that the data reduction rate η becomes less than one.

2.3. Decomposition to Frequency Bands by DWT

Applying the DWT horizontally and vertically to the video signal X, four kinds of frequency band components $\{LL1, HL1, LH1, HH1\}$ are produced. Repeating this procedure to the lowest frequency band $LL1$, other four bands $\{HH2, HL2, LH2, LL2\}$ are generated. As a result of n time octave decomposition, 4+3(n-1) kinds of frequency bands are generated from the input signal. We denote it by

$$\begin{aligned} &DWT[X] = F_r \\ &F_r = \{HH1, HL1, LH1, HH2, HL2, LH2, \cdots, LLn\} \end{aligned} \tag{3}$$

where F_r is the universal set of the frequency band components. In this report, we determine the family;

$$\begin{aligned} f_r \in \{ sLL \mid\ & sLL = HHm + HLm + LHm + \cdots + HHn \\ & + HLn + LHn + LLn,\ \ s = m-1,\ \ m \in \{2,3,\cdots,n-1\} \} \end{aligned} \tag{4}$$

to be included into the 1st layer for ME in the remote robot.

2.4. Decomposition to Bit Planes by BPD

Value of the frequency band signals are furthermore decomposed into bit planes by the BPD. We denote this as

$$\begin{aligned} BPD[F_r] &= F_r \times B_p \\ B_p &= \{8th, 7th, \cdots, 2nd, 1st\} \end{aligned} \tag{5}$$

where "1*st*" means the most significant bit (MSB) plane. We determine one of the families;

$$b_p \in \left\{ q \,\middle|\, q = \sum_{Q=1}^{q} Qth,\ \ q \in \{8,7,\cdots,2,1\} \right\} \tag{6}$$

to be included into S_R for each f_r, namely $f_r \times b_p$.

2.5. Extraction of the Components

The optimum set S_H for human communication is defined as the direct set $f_r \times b_p$ which minimizes its data volume $DS[f_r \times b_p]$ under the image quality θ required by a person for browsing. It is denoted by

$$S_H = \arg\min_{f_r \times b_p} DS[f_r \times b_p] \ \ subject\ to\ \ SN[f_r \times b_p] > \theta \tag{7}$$

where $SN[f_r \times b_p]$ denotes the L^2 norm metric to evaluate quality of the image reconstructed from $f_r \times b_p$.

On the other hand, in the proposed method, the optimum set S_R for the robot communication is defined as the direct set $f_r \times b_p$ which minimizes its data volume $DS[f_r \times b_p]$ under tolerable amount of the motion estimation error σ. It is denoted by

$$S_R = \arg\min_{f_r \times b_p} DS[f_r \times b_p] \ \ subject\ to\ \ MV[f_r \times b_p] < \sigma \tag{8}$$

where $MV[f_r \times b_p]$ represents error amount of the motion estimation from the video reconstructed by the components $f_r \times b_p$. In this report, we use the standard deviation of the motion estimation error.

3. MOTION ESTIMATION

In the experiment in this report, we evaluate the cross correlation (CC) [4], the sum of squared difference (SSD) [5] and the phase only correlation (POC) [6] as the motion estimation. The motion vector *mv* is estimated by

$$\begin{aligned} mv_i &= \arg \max_{\mathbf{m}} E_i, \\ i &\in \{CC, SSD, POC\}, \quad \mathbf{m} = (m_1, m_2) \end{aligned} \tag{9}$$

where pixel values of two frames at location $\boldsymbol{n}=(n_1,n_2)$ are denoted by $x_1(\boldsymbol{n})$ and $x_2(\boldsymbol{n})$. Denoting the forward transform and the backward transform of the discrete Fourier transform by $F[\ ]$ and $F^{-1}[\ ]$ respectively, the cost function E_i in eq.(9) is defined by the following equations.

(a) CC (cross correlation)

$$E_{CC} = F^{-1}\left[X_1(\mathbf{w})\overline{X_2(\mathbf{w})} \right] \tag{10}$$

(b) SSD (sum of squared difference)

$$E_{SSD} = -\sum_{\mathbf{n}} \{x_1(\mathbf{n}) - x_2(\mathbf{n}-\mathbf{m})\}^2 \tag{11}$$

(c) POC (phase only correlation)

$$E_{POC} = F^{-1}\left[\frac{X_1(\mathbf{w})\overline{X_2(\mathbf{w})}}{\left| X_1(\mathbf{w})\overline{X_2(\mathbf{w})} \right|} \right] \tag{12}$$

where

$$\begin{cases} X_q(\mathbf{w}) = F\left[x_q(\mathbf{n}) \right] = \sum_{\mathbf{n}} x_q(\mathbf{n}) W_N^{-\mathbf{w}\cdot\mathbf{n}} \\ \hat{x}_q(\mathbf{m}) = F^{-1}\left[X_q(\mathbf{w}) \right] = \frac{1}{N} \sum_{\mathbf{w}} X_q(\mathbf{w}) W_N^{\mathbf{w}\cdot\mathbf{m}} \end{cases} \tag{13}$$

$$W_N = e^{j2\pi/N}, \ \ q \in \{1,2\}.$$

4. SIMULATION RESULTS

In the experiment, two regions of 256×256 [pixel] at randomly different locations are extracted from the image of 480×360 [pixel] in Fig.1(b). The motion vector is estimated by eq.(10) or (11) or (12). The estimation is performed 300 times and the estimation error is evaluated.

4.1. Motion Estimation Error

Table 1 summarizes the standard deviation σ of the motion estimation error at various combinations of f_r and b_p. In case of the CC, it is observed that σ=0 at any b_p for f_r=*all* and 1*LL*. This means that only 1 bit plane and 1*LL* band is enough for precise motion estimation with σ=0. Among combinations which satisfy σ=0, the combination $f_r \times b_p$=1*LL*×1 has the minimum data volume. As a result, it is found to be the optimum components as the set S_R for the remote robot.

The POC tends to be less robust to the distortion in this experiment comparing to the SSD and the CC.

4.2. Data Compression Rate

Fig.5 illustrates the rate distortion curve of the CC. When θ=45.9 [dB] is required by a person for browsing, all the frequency bands and the bit planes must be transmitted. On the other hand, when σ=0 is required for the motion estimation, the minimum components are $f_r \times b_p$=1*LL*×1 and θ=4.64 [dB]. In this case, the data volume is reduced to η=0.712 [%] as indicated in table 2. A person can browse the image in Fig.6(a), however the image in Fig.6(b) has enough quality for the remote robot.

It is confirmed that the proposed method can reduce the data volume to 0.7～1.3 [%], under a given requirement of a function for communications between robots.

Table 1 Motion estimation error σ [pixel].

CC		f_r (frequency band)				
		all	1LL	2LL	3LL	4LL
b_p (bit plane)	8	0	0	0.07	0.48	3.71
	7	0	0	0.07	0.48	3.77
	6	0	0	0.07	0.48	3.72
	5	0	0	0.08	0.48	3.68
	4	0	0	0.09	0.49	3.62
	3	0	0	0.12	0.54	4.11
	2	0	0	0.16	0.62	4.11
	1	0	0	0.25	0.67	3.90

SSD		f_r (frequency band)				
		all	1LL	2LL	3LL	4LL
b_p (bit plane)	8	0	0	0	0.34	1.20
	7	0	0	0	0.34	1.19
	6	0	0	0	0.35	1.22
	5	0	0	0	0.37	1.25
	4	0	0	0	0.34	1.15
	3	0	0	0	0.49	1.66
	2	0	0	0.08	0.63	2.70
	1	0	0	0.27	0.74	2.89

POC		f_r (frequency band)				
		all	1LL	2LL	3LL	4LL
b_p (bit plane)	8	0	0.47	1.44	5.92	19.06
	7	0	0.47	1.44	5.91	19.06
	6	0	0.47	1.44	5.91	18.99
	5	0	0.47	1.44	5.92	19.20
	4	0	0.47	1.44	5.92	19.74
	3	0	0.47	1.44	6.07	20.85
	2	0	0.47	1.44	5.98	61.62
	1	0	0.47	1.44	6.04	110.50

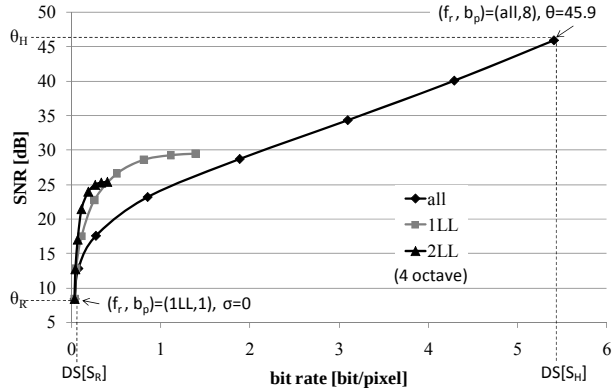
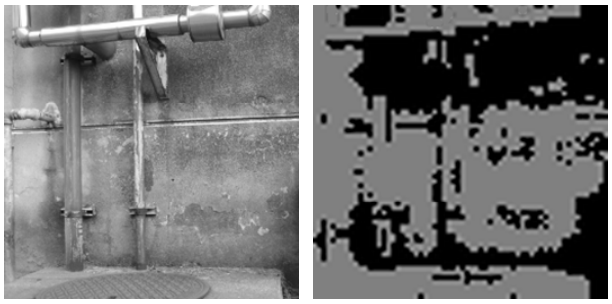


Fig.5 Rate distortion curve of the CC.



(a) Existing method. (b) Proposed method.

Fig.6 Reconstructed images for ME at the remote robot.

Table 2 Data reduction rate η [%].

CC		SNR θ [dB]		
		45.9	40.1	34.3
error σ [pixel]	0	0.712	0.896	1.243
	0.5	0.697	0.879	1.219
	1	0.696	0.877	1.217

SSD		SNR θ [dB]		
		45.9	40.1	34.3
error σ [pixel]	0	0.712	0.896	1.243
	0.5	0.697	0.879	1.219
	1	0.696	0.877	1.217

POC		SNR θ [dB]		
		45.9	40.1	34.3
error σ [pixel]	0	0.726	0.914	1.268
	0.5	0.712	0.896	1.243
	1	0.712	0.896	1.243

5. CONCLUSIONS

In this report, we proposed a functionally layered video coding for a robot vision network. Removing redundancy in the bit stream for a robot to robot communication, data volume to be transmitted is reduced to 0.7～1.3 [%].

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 20560351, 2008.

6. 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] 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.

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

[4] 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.

[5] 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. Robotcs and Automation, Vol. 9, Issue 1, pp. 14 - 35, Feb. 1993.

[6] 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.

[7] JTC1/ SC29, "Information technology - JPEG 2000 image coding system: Core coding system", ISO/IEC15444-1, 2004.

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

[9] M. Iwahashi, S. Udomsiri, Y. Imai, S. Muramatsu, "Water Level Detection for Functionally Layered Video Coding", IEEE ICIP, MP-P3.11, vol.II, pp.321-324, Sept. 2007.