Chinese Journal of Electronics
Vol.22, No.3, July 2013

Distributed Compressive Video Sensing with Adaptive Measurements Based on Structural Similarity*

LIU Zhuo[1], WANG Anhong[1], ZENG Bing[2], ZHANG Xue[1], BAI Huihui[3] and LI Zhihong[1]

(1.*Institute of Digital Media and Communication, Taiyuan University of Science and Technology, Taiyuan 030024, China*)

(2.*Hong Kong University of Science and Technology, Hong Kong SAR, China*)

(3.*Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China*)

Abstract — This paper presents a Distributed compressive video sensing scheme with Adaptive measurements (DCVS-AM). In this approach, the key frame in each Group of pictures (GOP) is coded by Compressive sensing (CS) with a fixed measurement rate; whereas other frames in the same GOP are compressed by an adaptive random projection in two stages, yielding the Adaptive compressive sensing (ACS) frames. The first stage uses a small and fixed measurement rate and recovers a coarse version. In the second stage, each coarse-version ACS-frame together with its proceeding and following key frames will go through a joint analysis at the decoder side and the analysis result – Structural similarity (SSIM) that is based on a motion-guided interpolation and calculated in a multilevel discrete wavelet transform domain – is sent back to the encoder side to facilitate a re-sampling of the ACS-frame with an adaptive measurement rate. Experimental results show that our proposed DCVS-AM consistently outperforms the state-of-the-art DCVS with a fixed measurement.

Key words — Distributed compressive video sensing (DCVS), Adaptive measurement rate, Structural similarity (SSIM), Discrete wavelet transform (DWT).

I. Introduction

One well-known feature of conventional video coding systems, such as MPEG and H.26x, is that they are highly asymmetric, *i.e.*, the encoder can be 5–10 times more complex than the decoder. In practice, such an asymmetric topology is very suitable to broadcasting and streaming applications where each source video is compressed only once (at the server side) but decoded many times (at the user side). In recent years, however, an increasing demand for the dual scenario (*i.e.*, the encoding is significantly less complex than the decoding) has emerged in up-link communications of low-power video capturing (*via* mobile cameras, wireless sensor network, *etc.*), where the computing power at the video-capturing end is highly limited.

Distributed video coding (DVC)[1], built on the Splepian-Wolf and Wyner-Ziv distributed source coding theories[2−4], is a framework developed to encode the highly-correlated video frames independently but decode them jointly. This framework has successfully shifted the computationally intensive operations (such as motion estimation/compensation and intra-prediction) to the decoder side, thus offering a good solution to the aforementioned scenario. In this paper, we follow the idea presented recently in Refs.[5, 6] to implement DVC through the Compressive sensing (CS) theory[7−9], leading to the Distributed compressive video sensing (DCSV) framework. In this framework, each source video frame is compressed independently by a number of random sampling operations (each being a simple and random linear projection) so as to keep the simplicity at the encoder side. On the other hand, motion analysis will be conducted at the decoder side, leading to a joint and more complicated decoding to deliver a higher performance.

Compared with the existing works in Refs.[5, 6], our contributions in this paper are summarized as follows.

• The existing DCVS schemes employ a fixed measurement (or sampling) rate to all frames, which ignores variations in the temporal correlation in a video sequence. In this paper, we propose a DCVS scheme with Adaptive measurements (DCVS-AM) over different frames so as to produce a better coding performance.

• The actual measurement rate for each frame is determined in our paper according to the popular Structural similarity (SSIM) metric – an objective assessment of image quality

*Manuscript Received July 2012; Accepted Sept. 2012. This work is supported in part by Sino-Singapore Joint Research Projet (No.2010DFA11010), the National Natural Science Foundation of China (No.61073142, No.61272262, No.61210006, No.61272051), the Doctor Startup Foundation of TYUST (No.20092011), the International Cooperative Program of Shanxi Province (No.2011081055), the Shanxi Provincial Foundation for Leaders of Disciplines in Science (No.20111022), Shanxi Province Talent Introduction and Development Fund (2011), Shanxi Provincial Natural Science Foundation (No.2012011014-3), the College Students Innovative Program of Taiyuan (No.120164077).

– that is computed at the decoder side in a multilevel Discrete wavelet transform (DWT) domain.

II. The DCVS Framework

The DCVS framework proposed in Ref.[5] is shown in Fig.1. A source video sequence is divided into several GOPs. Each GOP contains a key frame, followed by some other frames (named as the CS-frames). Each frame $\boldsymbol{p}_t$ (of size $N \times N$, t denoting the time) is first converted into a 1-D vector $\boldsymbol{x}_t$ (with height N^2) and then compressed *via* a CS-process as:

$$\boldsymbol{y}_t = \boldsymbol{\Phi}\boldsymbol{x}_t \tag{1}$$

where $\boldsymbol{y}_t$ is output vector after performing M_t measurements and $\boldsymbol{\Phi}$ represents the $M_t \times N^2$ measurement (or sampling) matrix. The measurement rate for $\boldsymbol{x}_t$ is denoted as $R_t = M_t/N^2$.

Compression is achieved through the random sampling shown in Eq.(1) due to the fact $M_t \ll N^2$. The CS theory[7−9] tells that M_t can be determined as $M_t = O(K \log N^2)$ for a K-sparse frame $\boldsymbol{x}_t$. It can be seen from Eq.(1) that the encoding process is extremely simple; while the sampling matrix $\boldsymbol{\Phi}$ is often generated through orthonormal i.i.d. Gaussians. In practice, one can choose to perform CS on a frame-by-frame basis, or block-by-block basis (in order to avoid maintaining a too big $\boldsymbol{\Phi}$ and reduce the complexity of the corresponding reconstruction at the same time[10,11]).

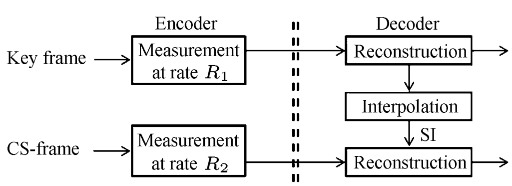


Fig. 1. The DCVS framework

According to Ref.[5], the key frame and all CS-frames in a GOP go through a regular CS-sampling, but with different measurement rates: R_1 (for the key frame) is greater than R_2 (for each CS-frame). An independent reconstruction is first carried out at the decoder side for all key frames. The reconstruction for a CS-frame is however much more complicated, in which both the proceeding and following key frames (reconstructed) will be used. More specifically, motion analysis (between two reconstructed key frames) will be conducted to derive a motion- guided interpolation so as to build a good target for each CS-frame, which is called the Side information (SI) in Ref.[5].

A very similar DCVS scheme has been proposed independently in Ref.[6] in which the key frame is coded by a conventional video coding standard (such as MPEG or H.26x) and a different reconstructing mechanism is used for each CS-frame.

III. Distributed Compressive Video Sensing with Adaptive Measurements (DCVS-AM)

The DCVS schemes proposed in Refs.[5] and [6] employ a fixed measurement rate for all CS-frames in a GOP. Apparently, they have ignored variations in the cross-frame correlation (*i.e.*, temporal correlation) within a video sequence. According to the Joint sparsity model (JSM)[12], the measurement rate for a CS-frame can be made smaller when the correlation between it and its reference (*e.g.*, the aforementioned target frame that is obtained from two key frames through interpolation) is larger. To determine an appropriate measurement rate for each CS-frame, one needs to carry out some analysis on the current CS-frame and its reference. Nevertheless, one must note that this analysis should not be done at the encoder side – it would otherwise defeat the purpose of maintaining a low-complexity at the encoder side.

In this paper, we propose to do such analysis at the decoder side. To this end, we employ the popular Structural similarity (SSIM) metric[13] to assess the correlation between two frames. To be more reliable, we follow the approach proposed in Ref.[14] to calculate the SSIM value in a multilevel Discrete wavelet transform (DWT) domain as follows (referring to Fig.2):

• The same multilevel DWT is applied to two images $\boldsymbol{X}$ and $\boldsymbol{Y}$ that are involved in the SSIM calculating procedure.

• One SSIM value is calculated separately within each frequency band.

• The final SSIM output, denoted as $SSIM_{\rm DWT}$, is obtained through a weighted sum of SSIM values in all frequency bands (using all w_ls that are given in Ref.[14]):

$$SSIM_{\rm DWT}(X,Y) = \frac{\sum_{l=1}^{L} \omega_l \cdot SSIM(X_l, Y_l)}{\sum_{l=1}^{L} \omega_l} \tag{2}$$

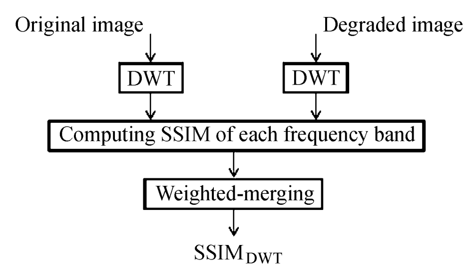


Fig. 2. Flowchart of the $SSIM_{\rm DWT}$ calculation

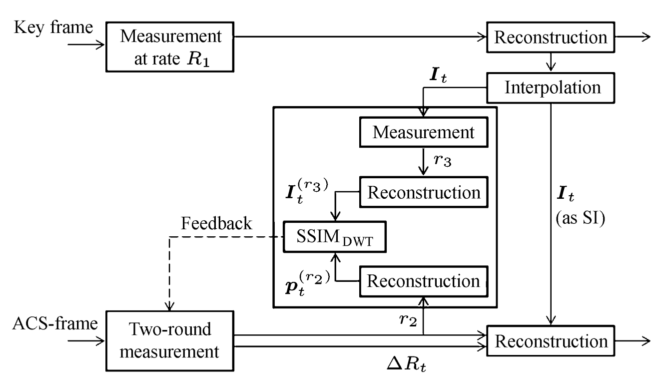


Fig. 3. Our proposed DCVS-AM system

Now, let's present our DCVS-AM scheme. As shown in Fig.3, it consists of an encoder with low-complexity and a decoder with high-complexity. When compared to the original DCVS scheme, a major change happens at the CS-sampling as well as the decoding process of each CS-frame. In particular, the latter change makes the decoder in the current framework even more complicated.

As depicted in Fig.3, each CS-frame $\boldsymbol{p}_t$ goes through an initial CS-sampling at rate r_2 (which is usually much smaller than R_1 – the rate used for the key frame). Then, an independent reconstruction is carried out at the decoder side to obtain a coarse version $\boldsymbol{p}_t^{(r2)}$.

Meanwhile, the interpolated frame $\boldsymbol{I}_t$ (through two reconstructed key frames) goes through a CS-process at rate r_3 and then the reconstructed frame $\boldsymbol{I}_t^{(r3)}$ is obtained. Finally, two images $\boldsymbol{p}_t^{(r2)}$ and $\boldsymbol{I}_t^{(r3)}$ are brought into the calculation of $SSIM_{DWT}$. Based on the $SSIM_{DWT}$ value, one can determine the extra sampling rate:

$$\Delta R_t = \begin{cases} \Delta R_1, & \text{if } 0 < SSIM_{DWT} < T_1 \\ \Delta R_2, & \text{if } T_1 \le SSIM_{DWT} < T_2 \\ \vdots & \\ \Delta R_n, & \text{if } T_{n-1} \le SSIM_{DWT} < 1 \end{cases} \tag{3}$$

where multiple thresholds T_is will be determined later on by experiments.

In practice, a simple law will be put in force: $\Delta R_1 \ge \Delta R_2 \ge \cdots \ge \Delta R_n$. The required extra rate ΔR_t is sent back to the encoder, assuming that a feedback channel is available (which is very feasible in the DVC scenario), so as to facilitate the second round sampling. Because of the time-varying nature of ΔR_t, the corresponding frames are called adaptive CS (ACS) frames.

The reason that we implement an extra CS-sampling (at rate r_3) on the interpolated frame $\boldsymbol{I}_t$ is to try to equalize two images $\boldsymbol{p}_t^{(r2)}$ and $\boldsymbol{I}_t^{(r3)}$ in terms of their quality level so as to yield a reliable SSIM value. To reach this goal, a natural choice is to let $r_2 = r_3$. Some experimental results will be presented in the next section to confirm this choice.

After the decoder receives all measurements from two rounds of CS-sampling (the resulted rate is $r_2 + \Delta R_t = R_2$), the final version of an ACS-frame will be reconstructed with aid of the side information, *i.e.*, $\boldsymbol{I}_t$, that is derived from the motion- guided interpolation (based on two key frames). Here, SI aids the reconstruction in two aspects: it is applied as the stopping criterion to speed up the reconstruction and simultaneously acts as the initialization of the iterative reconstructing algorithm to improve the performance, see Ref.[5] for the details.

IV. Experimental Results

All experimental results presented in this section are obtained from the frame-based processing over three CIF video sequences (150 frames totally in each sequence, luminance only): Coastguard, Foreman, and Mother-Daughter.

In the first set of experimental results, we assume that GOP size=2 and study how the $SSIM_{DWT}$ varies with the choice of r_2 and r_3. To this end, we considered three cases (for each sequence): $r_2 = r_3 = 0.1$ (equal but low rate); $r_2 = r_3 = 0.2$ (equal but high rate); and $r_2 = 0.1, r_3 = 0.2$ (unequal rates); whereas each key frame is sampled with $R_1 = 0.7$. Under this setup, we decompose two frames $\boldsymbol{p}_t^{(r2)}$ and $\boldsymbol{I}_t^{(r3)}$ (see Fig.3) through the 5-level DWT with Daubechies 9/7 filters and then calculate the corresponding $SSIM_{DWT}$. Notice that three high-frequency sub-bands LH, HL, and HH at each level are combined together so that $L = 6$ when using Eq.(2).

The results are shown in Fig.4. It is clear from these results that $SSIM_{DWT}$ drops significantly when r_2 and r_3 are unequal. Similar results have also been obtained when GOP size=3. Consequently, we always choose $r_2 = r_3$ in the following experiments.

From now on, we assume GOP size=3 and choose $R_1 = 0.7$ and $r_2 = r_3 = 0.1$, respectively. Since there is no difference on each key frame between the existing DCVS scheme and our DCVS-AM scheme, the following comparison focuses on the non-key frames.

• We implement our DCVS-AM scheme first, in which the extra rate ΔR_t is determined according to Table 1. We bookkeep the total rate $R_2 = r_2 + \Delta R_t$ consumed in sampling each

Table 1. Measurement rates assigned in our DCVS-AM

Sequences	$SSIM_{DWT}$	Adaptive measurement rate				
		Q1	Q2	Q3	Q4	Q5
Coastguard	[0,0.6]	0.15	0.3	0.4	0.5	0.6
	(0.6, 0.66]	0.05	0.2	0.3	0.4	0.5
	(0.66, 0.7]	0	0.1	0.2	0.3	0.4
	(0.7, 1]	0	0	0	0.1	0.1
Foreman	[0, 0.6]	0.1	0.25	0.35	0.5	0.55
	(0.6, 0.7]	0.05	0.2	0.3	0.4	0.45
	(0.7, 1]	0	0	0	0	0.1
Mother and daughter	[0, 0.89]	0.15	0.25	0.35	0.45	0.55
	(0.89, 0.92]	0.05	0.2	0.3	0.4	0.45
	(0.92, 1]	0	0	0	0	0.1

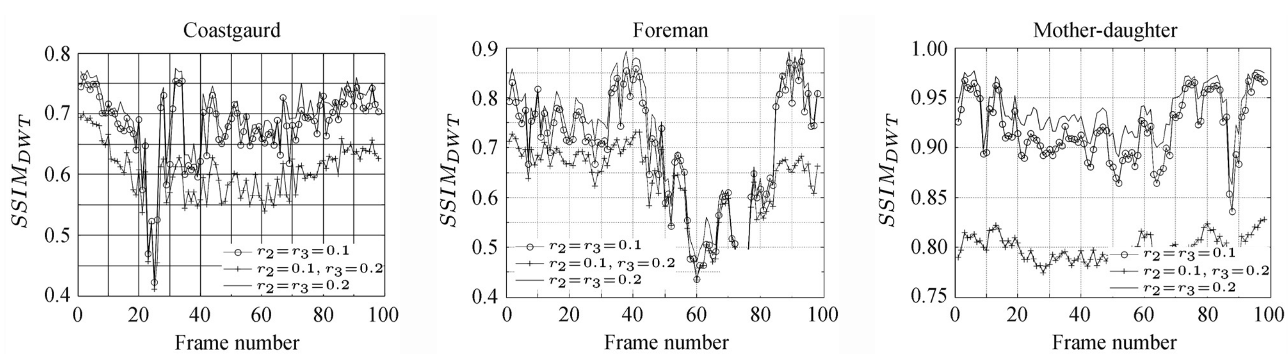


Fig. 4. $SSIM_{DWT}$ values under different CS-sampling rates r_2 and r_3

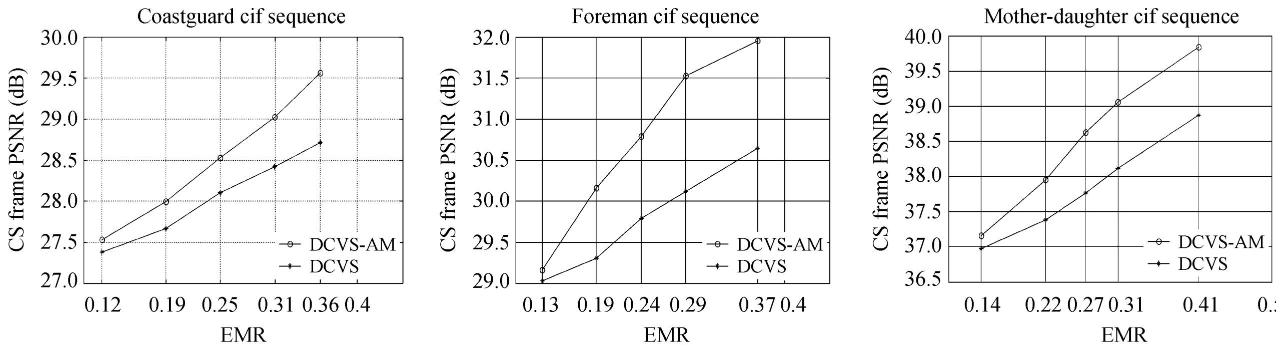


Fig. 5. Comparisons of our DCVS-AM and the fixed-rate DCVS

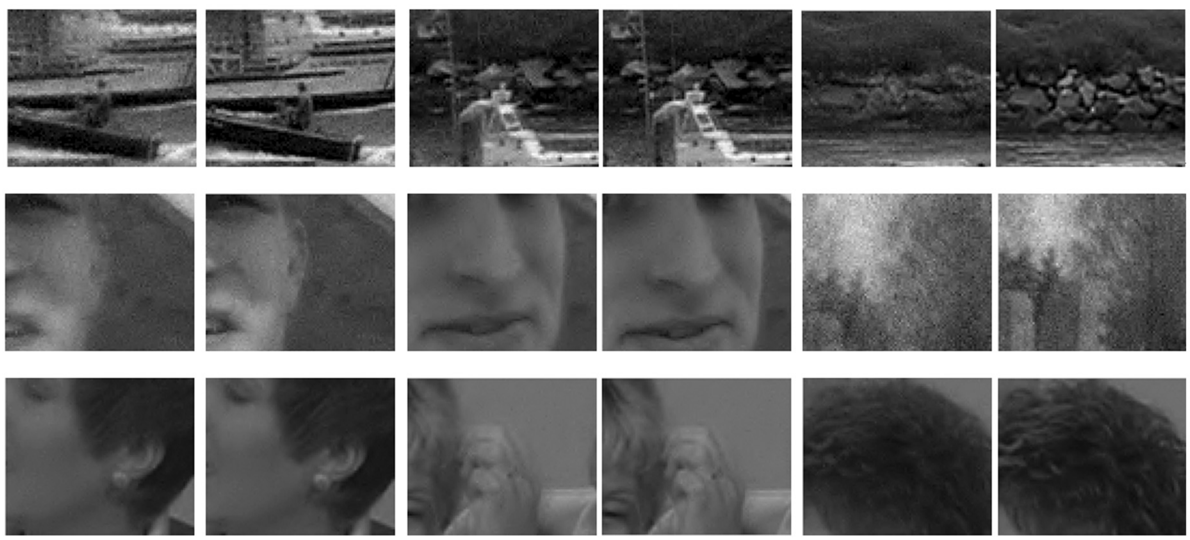


Fig. 6. Some reconstructed frames by DCVS (left) and DCVS-AM (right) at the same EMR: A side-by-side visual comparison

ACS-frame. We take average over all ACS-frames to obtain an Equivalent measurement rate (EMR).

• According to the obtained EMR, we then implement the DCVS scheme proposed in Ref.[5] (*i.e.*, each CS-frame is sampled at a fixed rate exactly equal to EMR) to facilitate a fair comparison.

As shown in Table 1, we have considered 5 quality levels (Q1–Q5) for each non-key frame according to SSIM_{DWT} calculated at the decoder side, *i.e.*, the measurement rate listed in Table 1 is assigned to ΔR_t. Fig.5 shows the comparison of our DCVS-AM and the fixed-rate DCVS[5] at these five quality levels. On average, our DCVS-AM achieves about 1dB PSNR gain over the DCVS with a fixed measurement rate.

For a visual comparison, we show in Fig.6 some reconstructed frames by DCVS and DCVS-AM at the same EMR. It is obvious that our DCVS-AM achieves better visual quality than DCVS due to the consideration of diversity of temporal correlation in video sequences.

V. Conclusions

We introduced in this paper a distributed compressive video sensing scheme in which adaptive measurement rates are applied on different frames. For each non-key frame in a GOP, the CS-sampling is implemented in two rounds: the sampling rate in the first round is fixed at a low level; whereas the rate in the second round is determined adaptively according to a SSIM- based analysis that involves the current reconstructed frame (in the first round) and two neighboring key frames (the proceeding and the following).

Experimental results demonstrated that the proposed DCVS-AM clearly outperforms the existing DCVS schemes with a fixed measurement rate. Since the encoder in such a DCVS system just consists of some random measurements, the nature of maintaining a low-complexity encoding is well preserved, which makes it very suitable for low-power mobile video capturing, such as mobile camera and wireless sensor networks.

References

[1] B. Girod, A. Aaron, S. Rane S and D. Rebollo-Monedero, "Distributed video coding", *Proc. of the IEEE*, Vol.93, No.1, pp.71–83, 2005.

[2] J.D. Slepian and J.K. Wolf, "Noiseless coding of correlated information sources", *IEEE Trans. Inf. Theory*, Vol.IT-19, No.4, pp.471–480, 1973.

[3] A.D. Wyner, "Recent results in the Shannon theory", *IEEE Trans. Inf. Theory*, Vol.IT-20, No.1, pp.2–10, 1974.

[4] S.S. Pradhan and K. Ramchandran, "Distributed source coding

using syndromes (DISCUS): Design and construction", *IEEE Data Compression Conf.*, Vol.49, No.3, pp.158–167, 1999.

[5] L.W. Kang and C.S. Lu, "Distributed compressive video sensing", *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Taibei, pp.1169–1172, 2009.

[6] T.T. Do, Y. Chen, D.T. Nguyen, N. Nguyen, L. Gan and T.D. Tran, "Distributed compressed video sensing", *IEEE Int. Conf. Image Processing*, Cairo, Egypt, pp.1393–1396, 2009.

[7] D.L. Donoho, "Compressed sensing", *IEEE Trans. Inf. Theory*, Vol.52, No.4, pp.1289–1306, 2006.

[8] E.J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?", *IEEE Trans. Inf. Theory*, Vol.52, No.12, pp.5406–5425, 2006.

[9] E.J. Candes and M.B. Wakin, "An introduction to compressive sampling", *IEEE Signal Processing Magazine*, Vol.25, No.2, pp.21–30, 2008.

[10] L. Gan, "Block compressed sensing of natural images", *Int. Conf. Digital Signal Processing*, Cardiff, pp.403–406, 2007.

[11] S. Mun and J.E. Fowler, "Block compressed sensing of images using directional trans-forms", *IEEE Int. Conf. Image Processing*, Cairo, pp.3021–3024, 2009.

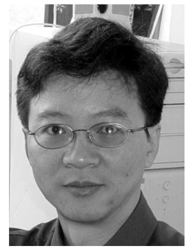
[12] M.F. Duarte, M.B. Wakin, D. Baron and R.G. Baraniuk, "Universal distributed sensing *via* random projections", *IEEE Int. Conf. Inf. Process. in Sensor Networks*, pp.177–185, 2006.

[13] Z. Wang, A.C. Bovik, "Image quality assessment: From error visibility to structural similarity", *IEEE Trans. on Image Processing*, Vol.13, No.4, pp.600–612, 2004.

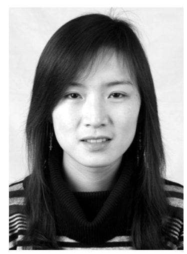
[14] C.L. Yang and W.R. Gao, "Research on image quality assessment in wavelet domain based on structural similarity", *Acta Electronica Sinica*, Vol.37, No.4, pp.845–849, 2009.

LIU Zhuo was born in Hubei Province in 1987. He received the B.S. degree from Hubei Engineering University. Now he is a postgraduate in school of Taiyuan University of Science and Technology. His current research focuses on compressive sensing. (Email: wojiushiliuzhuo@126.com)

WANG Anhong (corresponding author) was born in Shanxi Province in 1972. She received B.E and M.E. degrees from Taiyuan University of Science and Technology (TYUST) in 1994 and 2002 respectively, and Ph.D. degree from Institute of Information Science, Beijing Jiaotong University in 2009. She became an associate professor with TYUST in 2005 and became a professor in 2009. She is now the director of Institute of Digital Media and Communication, Taiyuan University of Science and Technology. Her research interest includes image/video coding, compressed sensing, and secret image sharing. She has published more than 30 papers. Now she is leading two national research projects from National Science Foundation of China. (Email: wah_ty@yahoo.com.cn)



ZENG Bing was born in Sichuan Province in 1963. He received the B.E. and M.E. degrees from the University of Electronic Science and Technology of China, Chengdu, in 1983 and 1986, respectively, and the Ph.D. degree from Tampere University of Technology, Tampere, Finland, in 1991, all in electrical engineering. He was a postdoctoral fellow with the University of Toronto and Concordia University in 1991 and 1992. Since 1993, he has been with the Hong Kong University of Science and Technology, where he is now an associate professor in the Department of Electronic and Computer Engineering. Professor Zeng is a member of IEEE. He served as an associate editor for the IEEE Trans. on Circuits and Systems for Video Technology (1995–1999). He has published over 100 referred papers in various journals and conference proceedings and holds 4 US patents.



ZHANG Xue was born in Shan dong Province in 1989. She received the B.S. degree from Linyi University. Now she is a postgraduate in Taiyuan University of Science and Technology. Her current research focuses on compressive sensing.

BAI Huihui was born in Shanxi Province in 1979. She is currently an associate professor in Institute of Information Science, Beijing Jiaotong University (BJTU). She received the B.E. and Ph.D. degrees in signal and information processing from Beijing Jiaotong University (BJTU) in 2001 and 2008 respectively. She has been engaged in R&D work in video coding technologies such as Multiple description video coding (MDC) and Distributed video coding (DVC). She has participated in the projects on distributed multiple description video coding from NSFC and 863 Program.

LI Zhihong was born in Shanxi Province in 1970. He is currently an associate professor in Taiyuan University of Science and Technology (TYUST). He received the B.E. degree in electronic information engineering from Taiyuan University of Science and Technology (TYUST) in 1994. His research interest includes compressed sensing and secret image sharing. He has participated in the projects on distributed video coding and now is leading one research project on image secret from Shanxi Natural Science Foundation.