International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.6, No.6 (2013), pp.1-12
http://dx.doi.org/10.14257/ijsip.2013.6.6.01

Power-aware Regions-of-Interest Computational Resource Allocation for Mobile Sign Language Video Encoding

Xiaolei Chen[1], Aihua Zhang[1] and Xinzhu Yang[2]

[1]*School of Electrical and Information Engineering*
Lanzhou University of Technology, Lanzhou, P. R. China
[2]*School of Mechanical Engineering*
Lanzhou Jiaotong University, Lanzhou, P. R. China

zhangah@lut.cn

Abstract

The primary challenge to design a mobile sign language communication system is overcoming the limitation of battery power. A scheme which allocates the computational resource of the encoder adaptive to available battery power and deaf people's visual system is proposed. In the scheme, encoding levels which determine number of reference frames and search range are adaptively selected according to the battery power and frame complexity at frame level. Then possible partition mode and quantization parameter are adaptively adjusted at the macro block (MB) level according to the relative priority of each MB. Experimental results show that the proposed algorithm obtains better peak-signal-noise-rate of face and hands that improves the intelligibility of sign language video. Encoding time can be saved by up to 89.4% compared to the standard H.264 encoder. Increased bit rate of the proposed algorithm is less than 5%, which is negligible.

Keywords: *power-aware, region-of-interest, resource allocation, sign language video*

1. Introduction

Sign language is a visual language used by deaf and hearing-impaired people to communicate. In the past, the main mobile communication way among deaf people is text messaging. Unfortunately, text messaging is much slower than signing, the speed of text conversation is limited by typing ability, and is at least 10 times slower than that of sign language. Furthermore, text messaging forces deaf users to communicate in spoken language as opposed to sign language. Therefore, for deaf people, text messaging is not an effective way of mobile communication. With the rapid development of mobile communication, availability of mobile devices with video cameras and emergence of the advanced video coding standard H.264, researchers proposed the mobile sign language communication technique [1]. This technique enables mobile devices to capture, encode, transmit, receive and decode sign language video, all in real-time, and greatly improves the mobile communication among deaf people.

H.264/AVC is one of the latest video coding standards. It provides a bit rate reduction of 50% as compared to MPEG-2 with similar subjective visual quality and thus is suitable for video communication between mobile devices. However, since H.264 video encoding is energy-demanding and mobile devices are powered by battery, the operational lifetime of mobile devices is short, mostly in the range of a few hours. This has become a challenge for technological progress in mobile sign language video communication.

ISSN: 2005-4254 IJSIP

According to the human visual system (HVS) research [2, 3], human vision is only able to focus on one or two areas in a frame, which is defined as region-of-interest (ROI), this is due to the limited capacity of human brain. Usually, HVS has different subjective attention to different parts of video and always pays more attention to ROI than others. The reported research demonstrates that for deaf people the most important region of sign language video is face, then is moving hands, and the last is background [4, 5]. This means for sign language video, face and hands are ROI, background is non-ROI. This phenomenon gives a chance to code all MB unequally: ROI has higher priority which can be allocated more computational resource; non-ROI has lower priority which can be allocated less computational resource. Liu *et al.*, [6] proposed a ROI based H.264 computational resource allocation scheme, in their scheme, several coding parameters including quantization parameter **(**QP), candidates for mode decision, number of reference frames (RF), accuracy of motion vectors and search range (SR) of motion estimation are adaptively adjusted at MB level according to the relative importance of each MB. It saves energy and guarantees fine quality in ROI. Similar other methods have been are proposed [7-9]. Summarizing, state-of-the-art ROI based computational resource allocation schemes offer a fixed low-power solution that may work well for short duration encodings of video, but do not perform well in case of long duration real-time encodings and ultimately are less power efficient. The researches [10,11] demonstrate that incorporating the third dimension of power consumption into conventional rate-distortion (R-D) analysis gives us one extra dimension of flexibility in computational resource allocation and allows us to achieve significant energy saving.

Based on the above analysis, with consideration of both deaf people's visual system and power consumption, in this paper we propose a novel adaptive power-aware regions-of-interest computational resource allocation for mobile sign language video encoding to treat the run-time changing scenarios of available energy budgets while keeping a good video quality in ROI. The proposed scheme has the following two steps: firstly, RF and SR are adaptively selected according to battery power and frame complexity at frame level. Then, possible partition modes and QP are adaptively adjusted at the MB level according to the relative priority of each MB. Experimental results show that the proposed algorithm greatly reduces computational complexity of H.264 and prolongs battery life while keeps fine quality in face and hands of sign language video.

This paper is organized as follows. In Section 2, the framework of whole encoding system is illustrated. The frame-level computational resource allocation with consideration of power consumption and frame complexity is illustrated is Section 3. In Section 4, the MB-level computational resource and bit allocation with consideration of deaf people's visual system is proposed. Section 5 shows the experimental results. Finally the conclusions are given in Section 6.

2. System architecture

The proposed adaptive power-aware regions-of-interest computational resource allocation scheme is depicted in Fig. 1. Four major steps are as follows:

Step1. Detect current battery power of mobile devices by calling the Windows API function. Detect ROI in sign language video using the method in subsection 4.1.

Step2. Determine RF and SR at frame-level according to battery power and frame complexity using the method in Section 3.

Step3. Determine possible partition modes and QP for different regions of sign language video at MB-level using the method in subsection 4.2.

Step4. Perform H.264 encoding using above determined parameters.

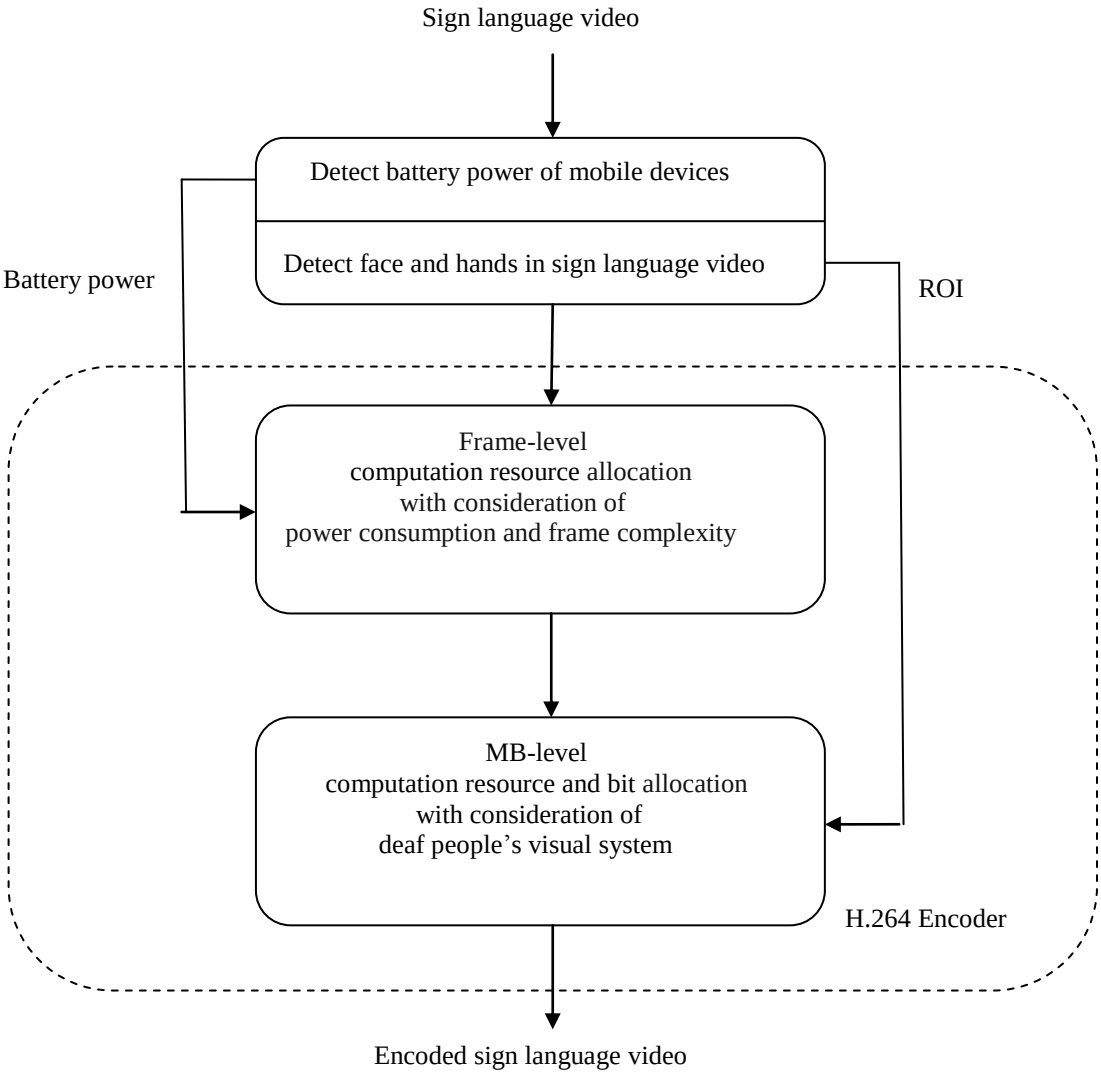


Figure 1. Encoding System Architecture

3. Frame-level Computational Resource Allocation with Consideration of Power Consumption and Frame Complexity

The number of RF and SR are two important parameters related to encoding time. More RF and larger SR achieve higher compression ratio, in expense of more encoding time. Therefore, at frame-level, our computational resource allocation scheme executes at different operational levels, each with different RF and SR.

For mobile devices, battery power decreases gradually with the working time. If battery power is sufficient, high computational power can be used to ensure video quality. If battery power is insufficient, low computational power can be used to reduce energy consumption, to ensure video encoding tasks can be finished before the energy is exhausted. Fig.2 illustrates the run-time transitions from one operational level to the other. Therefore, we choose battery

power (BP) as one factor to determine RF and SR.

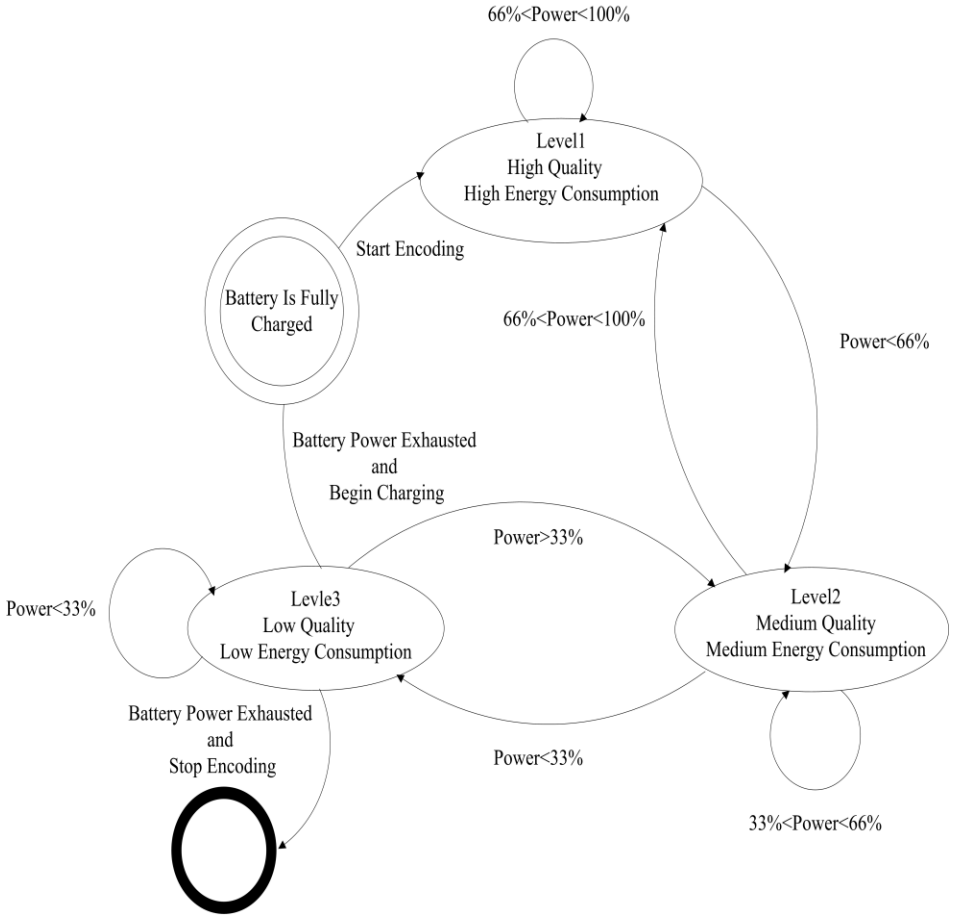


Figure 2. Flowchart of Run-time Transitions from One Operational Level to the Other

Frames with higher complexity consume more computational resource, and frames with lower complexity consume fewer computational resource. Based on this observation, we choose frame complexity as the other one factor to determine RF and SR. We estimate frame complexity (FC) derived from PSNR drop ratios [12].

Using BP and FC, we set RF and SR as follows:

$$RF = ceil(BP + FC * 0.65) \quad (1)$$

$$SR = \min((BP + 12 * (FC * 0.65)),32) \quad (2)$$

4. MB-level Computational Resource and Bit Allocation with Consideration of Deaf People's Visual System

H.264/AVC performs RD optimization with more flexible MB partition mode for Intra and Inter frames. The mode selection algorithm computes the best possible mode from a set of partition modes using Lagrangian method. However, RDO-based mode decision includes a

series of operations for each MB partition mode and requires high computational power. Many research works [13, 14] have been done to reduce the computation complexity of mode selection and different kinds of fast algorithms have been proposed. These fast algorithms regard all MBs as the same priority. However, for deaf people the most important region of sign language video is face, then is moving hands, and the last is background. This means for sign language video, face and hands is ROI, background is non-ROI. Therefore, spending so much computational resource on mode decision for MBs in background is not necessary. Furthermore, original rate controller of H.264 allocates same QP for all MBs, this can not improve the quality of ROI. In our encoder, the possible partition mode and QP are adaptively adjusted at MB level according to the relative priority of each MB. There are two major problems when the concept of ROI is used in computational resource and bit resource allocation with consideration of deaf people's visual system: the detection of ROI in sign language video, and the ROI based computational power and bits allocation.

4.1. Fast ROI detection for Sign Language Video

In the proposed MB-level computational resource and bit resource allocation algorithm, ROI detector is a pre-processor of the algorithm. Accurate face and hands detectors need a large number of computational resources. They are hardly implementable on battery powered mobile devices due to their high computational requirements. Moreover, high level of accuracy of face and hands detection is not required for MB-based video encoding. A rough face and hands information is enough to create the MB level ROI mask. Therefore, in this paper, we propose a fast face and hands detector in sign language video as shown in Figure 3. The difference between current frame and previous frame is computed to derive temporal information. Meanwhile, image pixels are classified as skin or non-skin to generate spatial information. Then the temporal information and spatial information are fused to generate a face and hands segmentation mask.

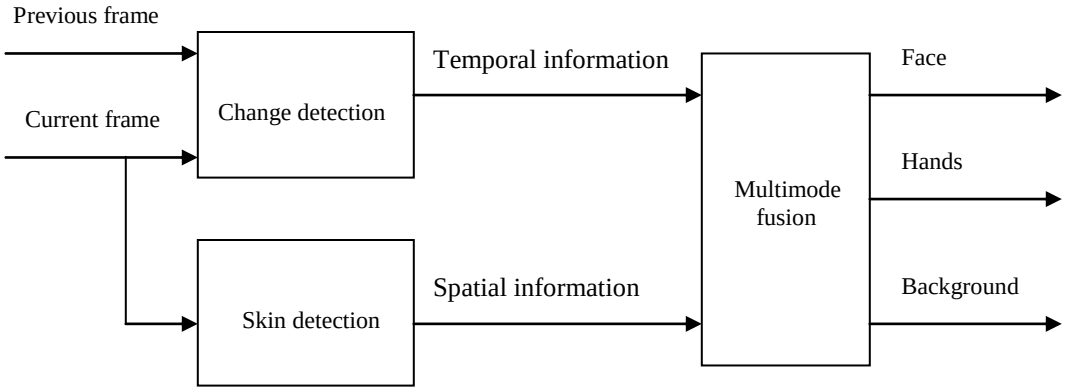


Figure 3. Fast ROI Detector for Sign Language Video

4.1.1. Change Detection: We derive temporal information of sign language video by a change detector. Change detection methods are generally computationally less expensive than motion estimation and optical-flow methods, and would therefore promote real-time segmentation.

Suppose $F(x,y,k)$ is current frame, $F(x,y,k-1)$ is previous frame, we can calculate the difference between the two frames as

$$DF_{k,k-1}(x,y)=\begin{cases}1, & |F(x,y,k)-F(x,y,k-1)|>Thr_k\\ 0, & else\end{cases} \quad (3)$$

All pixels in $DF_{k,k-1}(x,y)$ with value 1 are regarded as moving pixels, all pixels in $DF_{k,k-1}(x,y)$ with value 0 are regarded as still pixels. Thr_k is a threshold for $DF_{k,k-1}(x,y)$ given by [15]

$$Thr_k=\frac{1}{M\cdot N}\sum_{x=1}^{M}\sum_{y=1}^{N}|F(x,y,k)-F(x,y,k-1)| \quad (4)$$

where M and N denote the numbers of the row and column in a frame respectively.

4.1.2. Skin Detection: We derive spatial information of sign language video by segmenting a video sequence into skin (*i.e.*, face and hands) and non-skin (*i.e.*, background) regions. Skin detection is feasible because human skin has a color distribution that differs significantly, although not entirely, from those of the background objects. Since digital video is stored and coded in YC_bC_r color space and conversion from one color space to another is computationally expensive. Therefore, we choose YC_bC_r color space in our research to detect skin. In the YC_bC_r color space, Y is the luminance component and C_b and C_r are the chrominance components. We assign a pixel as skin if the C_b component is between 77 and 127 and the C_r component is between 133 and 173 as

$$p=\begin{cases}1 & if \quad C_b\in[77,127] \quad and \quad C_r\in[133,173]\\ 0 & else\end{cases} \quad (5)$$

where p is a binary parameter to represent whether the pixel is detected as skin or not (1: skin, 0: non-skin).

4.1.3 Generation of Face and Hands Detector: The pixel level change detection and skin detection information are fused to indicate the face and hands detection mask at the MB level by

$$I_{ROI}[i]=\sum_{x=1}^{16}\sum_{y=1}^{16}[DF_{k,k-1}(x,y)\cdot p(i,x,y)] \quad (6)$$

where i is the MB index, x,y are the pixel index within one MB.

4.2. ROI based Computational Power and Bits Allocation

After detection of face, hands and background, the mode decision with consideration of deaf and hearing-impaired people's visual attention is proposed and depicted in Figure 4. While coding background MB, {Skip, P16×16} is chosen ,while coding hands MB , {Skip, P16×16, P16×8, P8×16, P8×8} is selected, while coding face MB ,all partition modes are used to maintain good perceptual quality.

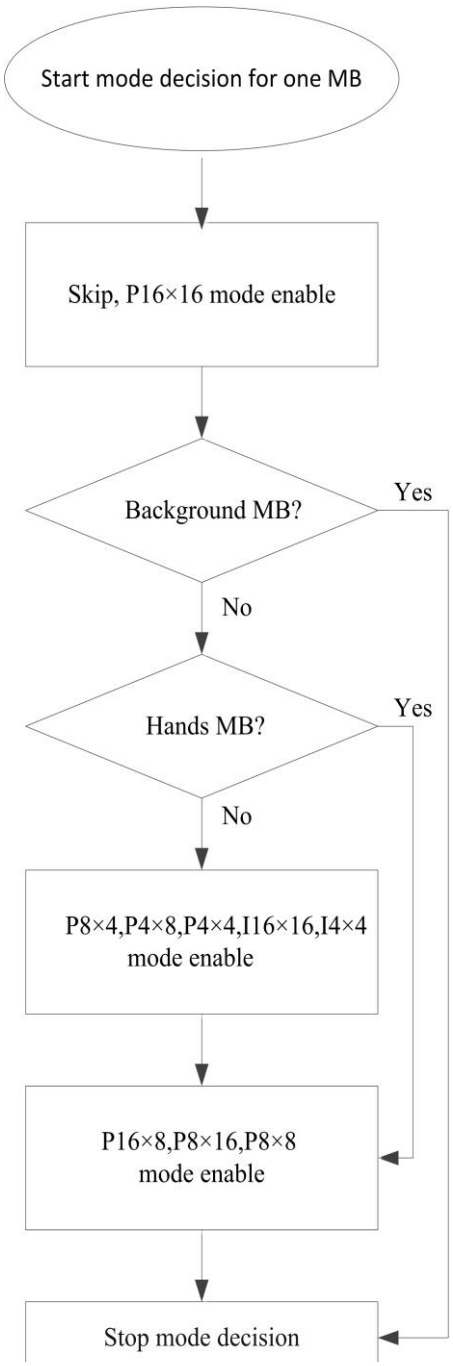


Figure 4. Flowchart of the Fast Mode Decision with Consideration of Deaf People's Visual Attention

Furthermore, in order to improve the quality of face and hands, original rate controller of H.264 is modified since the existing rate controller consider priorities of all MBs are the same with each other and allocates same QP for all MBs. In our encoder, the quantization parameter $QP_{i,MB(j)}$ for the i^{th} frame and j^{th} MB, is determined as follows:

$$QP_{i,MB(j)} = \begin{cases} \min(QP_i + floor(QP_i / a) + b, 51) & j \in MB(background) \\ QP_i - 5 & j \in MB(hands) \\ QP_i - 10 & j \in MB(face) \end{cases} \quad (7)$$

Where QP_i is the quantization parameter for the i^{th} frame determined by the native rate controller, and $QP_{i,MB(j)}$ is the refined QP for the i^{th} frame and j^{th} MB, a and b are constants with values 8 and 2 respectively. According to the equation (7), the modified rate controller can assign more bits to face and hands, fewer bits to background.

5. Experimental Results

The performance of the proposed encoding scheme is evaluated in this section. Two standard sign language test sequences Silent and Irene with QCIF (176×144) spatial resolution are carried out in our experiments. The simulation is implemented with JM16.2, in which the proposed computational power allocation methods are employed in our encoder. The comparison items include encoding time, PSNR_Y of face, hands and background, and bit rate. All the test sequences are intra-coded for the first frame (I frame) and followed with inter-coded frames (P frames). The coding parameters of original JM16.2 are: baseline Profile. RDO and CABAC are enabled. RF number is 3, SR is 32, prediction modes is {Skip, P16×16, P16×8, P8×16, P8×8, P8×4, P4×8, P4×4, I16×16, I4×4},QP is set to 28.The coding parameters of the proposed encoder are adaptively adjusted for our power-aware regions-of-interest computational resource allocation scheme. The experiment platform is a lithium battery powered laptop. The hardware configuration of the laptop is as follows: CPU 2.13GHz, memory 2G DDR. Table 1 gives the details of performance for JM16.2, literature [6] and three operational levels of our encoder for two standard sign language test sequences.

Table 1. Performance Comparison (100 frames, 30 fps)

Sequence	Method	Encoding time(s)	PSNR(dB)				Bit rate(kb/s)
			Face	Hands	Back-ground	Average	
Silent	JM16.2	409.835	35.69	35.31	37.22	37.02	82.53
	[6]	400.275	35.98	35.38	33.45	35.12	82.16
	Level1+ROI	392.947	36.38	35.73	33.90	34.73	79.41
	Level2+ROI	104.040	36.15	35.48	33.56	34.56	81.84
	Level3+ROI	43.600	35.99	35.40	33.30	34.17	86.08
Irene	JM16.2	407.637	36.32	36.04	40.13	37.71	125.74
	[6]	401.257	36.34	36.10	35.19	35.79	120.38
	Level1+ROI	397.335	36.94	36.58	35.51	35.62	113.72
	Level2+ROI	102.783	36.69	36.40	35.25	35.46	118.52
	Level3+ROI	43.076	36.36	36.20	35.08	35.15	124.94

For Silent sequence compared to JM our Level1+ROI, Level2+ROI, Level3+ROI achieve encoding time reduction of 4.12%, 74.6%, 89.1% at the cost 2.29 dB, 2.46 dB, 2.85dB loss in average PSNR respectively, meanwhile, the PSNR in face increases 0.69 dB, 0.46 dB, 0.30dB, the PSNR in hands increases 0.42 dB, 0.17 dB, 0.09dB, increased bit rate of Level3+ROI is 4.9%. Compared to [6], our Level1+ROI, Level2+ROI, Level3+ROI achieve encoding time reduction of 1.8%, 74.0%, 89.1% at the cost 0.39 dB, 0.56 dB, 0.95dB loss in average PSNR

respectively, meanwhile, the PSNR in face increases 0.40 dB, 0.17 dB, 0.01dB, the PSNR in hands increases 0.35 dB, 0.10 dB, 0.02dB, increased bit rate of Level3+ROI is 4.7%

For Irene sequence compared to JM our Level1+ROI, Level2+ROI, Level3+ROI achieve encoding time reduction of 2.53%, 74.8%, 89.4% at the cost 2.09 dB, 2.25dB, 2.56dB loss in average PSNR respectively, meanwhile, the PSNR in face increases 0.62 dB, 0.37 dB, 0.04dB, the PSNR in hands increases 0.54 dB, 0.36 dB, 0.16dB, increased bit rate of Level3+ROI is 2.5%. Compared to [6], our Level1+ROI, Level2+ROI, Level3+ROI achieve encoding time reduction of 0.97%,74.3%, 89.2% at the cost 0.17 dB, 0.33 dB, 0.64 dB loss in average PSNR respectively, meanwhile, the PSNR in face increases 0.60 dB, 0.35 dB, 0.02dB, the PSNR in hands increases 0.48, 0.30, 0.10dB, increased bit rate of Level3+ROI is 3.8%.

Since our method divides each frame into three parts and codes them unequally, further than PSNR, the subjective visual quality evaluation should be introduced. Figure 5 shows the comparative results. In the frames which are coded by the proposed scheme, more details are kept on face and hands, this enhances the intelligibility of sign language video.

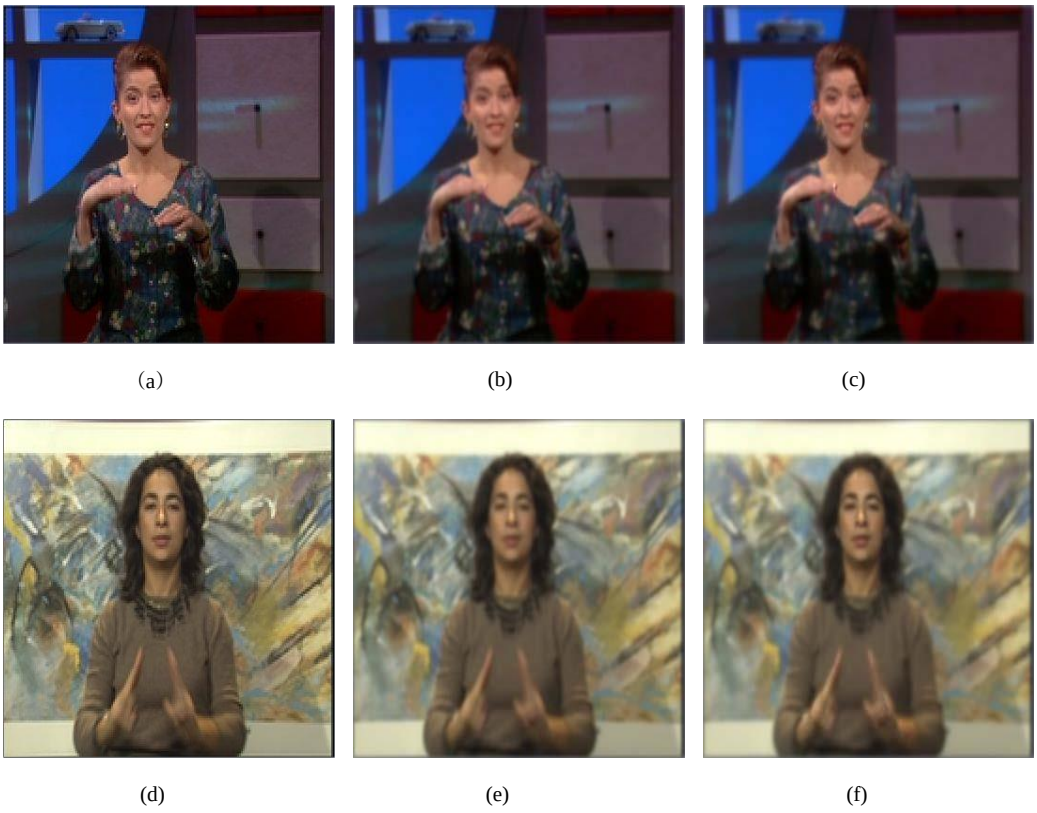


Figure 5. Example of Comparison of the Subjective Visual Quality. a Original 21th Frame of Irene. b the 21th Frame Encoded by JM16.2. c the 21th Frame Encoded by the Proposed Method (Level1+ROI).d Original 219th Frame of Silent.e the 219th Frame Encoded by JM16.2.f the 219th Frame Encoded by the Proposed Method. (Level1+ROI)

6. Conclusions

In this paper, we proposed a novel mobile sign language video encoding algorithm with consideration of both deaf people's visual system and power consumption. It consists of a frame-level computational resource allocation algorithm with consideration of power consumption and frame complexity, a MB-level computational resource allocation algorithm with consideration of deaf people's visual system. The experimental results show that the proposed computational resource allocation algorithm keeps fine quality in face and hands of sign language video, greatly reduces the computational complexity and prolongs battery life. It can be applied to meet the requirement of mobile sign language communication. Furthermore, the algorithm can be implemented in conjunction with other battery powered mobile multimedia applications without any impediments.

Acknowledgements

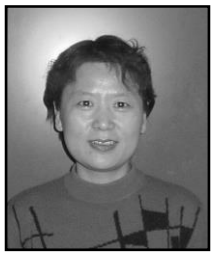
This work is supported by the National Natural Science Foundation of China under grant No. 61302116 and 61365003.It is also supported by the Natural Science Foundation of Gansu Province, China under grant No. 1212RJYA026 and 1212RJZA050.

References

[1] F. M. Ciaramello and S. S. Hemami, "A computational intelligibility model for assessment and compression of American sign language video", IEEE Transactions on Image Processing, vol. 20, **(2011)**, pp. 3014-3028.
[2] Y. M. Fang, W. S. Lin and B. S. Lee, "Bottom-Up Saliency Detection Model Based on Human Visual Sensitivity and Amplitude Spectrum", IEEE Transactions on Multimedia, vol. 14, **(2012)**, pp. 187-198.
[3] U. Engelke, H. Kaprykowsky and H. Zepernick, "Visual attention in quality assessment", IEEE Signal Processing Magazine, vol. 28, **(2011)**, pp. 50-59.
[4] A. Dimitris, C. Nishan and R. David, "A perceptually optimized video coding system for sign language communication at low bit rates", Signal Process: Image Communication, vol. 21, **(2006)**, pp. 531-549.
[5] B. Davide, V. Matteo and P. Francesco, "Prominent reflexive eye-movement orienting associated with deafness", Cognitive Neuroscience, vol. 3, **(2012)**, pp. 8-13.
[6] Y. Liu, Z. G. Li and C. S.Yeng, "Region-of-interest based H.264 encoding parameter allocation for low power video communication", IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, **(2008)**, pp. 134-139.
[7] Y. Y. Zheng, F. Zhou and X. Tian, "Lightweight content-adaptive coding in joint analyzing-encoding framework", IEEE Transactions on Consumer Electronics, vol. 54, **(2008)**, pp. 614-622.
[8] T. R. Zhang, L. Chen and M. H. Wang, "Multiple region-of-interest based H. 264 encoder with a detection architecture in macroblock level pipelining", IEICE Transactions on Electronics, vol. E94.C, **(2011)**, pp. 401-410.
[9] P. Y. Liu and K. B. Jia, "Research and optimization of low-complexity video encoding method based on visual perception", The Journal of Information Hiding and Multimedia Signal Processing, vol. 2, **(2011)**, pp. 217-226.
[10] Z. H. He, W. Cheng and X. Chen. "Energy minimization of portable video communication devices based on power-rate-distortion optimization", IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, **(2008)**, pp. 596-608.
[11] W. K. Li, S. L. Chun and Y. L. Chih, "Low-complexity video coding via power–rate–distortion optimization", Journal of Visual Communication and Image Representation, vol. 23, **(2012)**, pp. 569-585.
[12] M. Jiang and N. Ling, "On Enhancing H.264/AVC Video Rate Control by PSNR-Based Frame Complexity Estimation", IEEE Transactions on Consumer Electronics, vol. 51, **(2005)**, pp. 281-287.
[13] K. Seongwan, L. Jaeho and P. Daehyun, "Fast block size and mode decision algorithm for intra prediction in H.264/AVC", IEEE Transactions on Consumer Electronics, vol. 58, **(2012)**, pp. 654-660.
[14] K. C. Chiang, M. F. Wu and J. J. J. Shann. "Modification and implementation of an edge-based fast intra prediction mode decision algorithm for H.264/AVC high resolution real-time systems", Journal of Visual Communication and Image Representation, vol. 23, **(2012)**, pp. 245-253.
[15] Y. Sun, A. Ishfaq and D. D. Li, "Region based rate control and bit allocation for wireless video transmission", IEEE Transactions on Multimedia, vol. 8, pp. 1-10, **(2006)**.

Authors

Xiaolei Chen, Male, was born in Henan, China in 1979, He took the Bachelor's degree and the Master Science in Electronic Engineering at the Lanzhou University in 2003 and 2006, respectively. He is currently pursuing a Ph.D. in Electronic Engineering. His research interests include image, video processing and wireless communication.



Aihua Zhang, Female, was born in Hebei, China in 1964, She received her Ph.D. degrees from Xi'an Jiaotong University in 2005. Since 2004, she has been with School of Electrical and Information Engineering at Lanzhou University of Technology as a Full Professor. Her main research interests are biomedical signal processing.

Xinzhu Yang, Female, was born is Hunan, China in 1981, She took the Bachelor's degree and the Master Science in Electronic Engineering at the Lanzhou Jiaotong University in 2003 and 2010, respectively. Her main research interests are signal processing and education management.

