

Perceptually optimised sign language video coding based on eye tracking analysis

D. Agrafiotis, N. Canagarajah, D.R. Bull and M. Dye

A perceptually optimised approach to sign language video coding is presented. The proposed approach is based on the results (included) of an eye tracking study in the visual attention of sign language viewers. Results show reductions in bit rate of over 30% with very good subjective quality.

Introduction: Coding of image sequences will always result in some information being lost, especially at low bit rates. With sign languages being visual languages, good image quality is necessary for understanding. Information loss should be localised so that it does not significantly impair sign language comprehension.

In this Letter we describe a foveated approach to coding sign language sequences with H.264 at low bit rates. We base our proposal on the results of an eye tracking study that we have conducted which allows us to analyse the visual attention of sign language viewers. Foveated processing is applied prior to coding in order to produce a foveation map which drives the quantiser in the encoder. Our coding approach offers significant bit rate reductions in a way that is compatible with the visual attention patterns of deaf people, as these were recorded in the eye tracking study.

Eye tracking study: Eleven subjects took part in the experiments, including deaf people, hearing signers and hearing beginners in British Sign Language (BSL). The experiments involved watching four short narratives in BSL. The clips were displayed uncompressed in the CIF format (352 × 288, 4:2:0) at 25 frames per second (fps). The Eyelink eye tracking system was used to record the participants eye-gaze while watching the four clips. Analysis of the results [1], and mainly of the fixation location and duration (i.e. locus at which eye-gaze is directed) showed that sign language viewers, excluding the hearing beginners, seem to concentrate on the facial area and especially the mouth. Most of the participants never looked at the hands while only a few showed just a small tendency to look at the hands. In contrast, hearing beginners did look at the hands more frequently (mainly due to a lack of understanding). Fig. 1 summarises some of the results in terms of the vertical position y of the recorded fixation points for a number of subjects and for one of the clips.

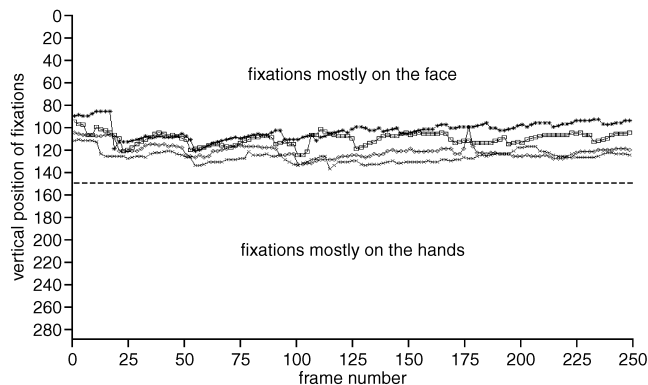


Fig. 1 Vertical position of fixation points for first 250 frames of eye-tracking clip-2 for four participants

—×— participant 2 —◇— participant 3
 —□— participant 5 —*— participant 8

Foveated processing: Foveated video compression [2, 3] aims to exploit the fall off in spatial resolution of the human visual system away from the point of fixation in order to reduce the bandwidth requirements of compressed video. A problem associated with foveated processing is the fact that the viewer's point of fixation has to be known, something that in practice requires real-time tracking of the viewer's eye-gaze. In our case (sign language viewers) and based on our study, the point of fixation lies almost always on the face, and specifically close to the mouth, thus removing the need to track the viewer's eye-gaze.

We have followed the local bandwidth approach as described in [3]. According to this method the video image is partitioned into eight

different regions based on their eccentricity (effectively distance from the centre of fixation) with the regions being constrained to be the union of disjoint macroblocks (MBs). The formula used to calculate the foveation regions together with suggested fitting parameters are given in [1] and [2]. Eccentricity e depends on viewing distance and is given by:

$$e = \tan^{-1} \left(\frac{d(x)}{Nv} \right) \quad (1)$$

where $d(x)$ is the Euclidean distance of pixel x from the fixation point, N is the width of the image and v is the viewing distance (in 100s of pixels) with all distances and co-ordinate measurements being normalised to the physical dimensions of pixels on the viewing screen. Foveated processing produces a map showing the region each MB belongs to for each frame. Fig. 2 shows one such map for $v=3$ alongside the corresponding frame of sequence 'FDD'. In [1] we have used such maps to pre-filter the MBs of each region with a lowpass filter, which had a lower cutoff frequency for MBs of higher eccentricity. In this work we use the foveation map to assign a different quantisation parameter (QP) to each MB with MBs lying in regions away from the point of fixation being allocated a higher QP.

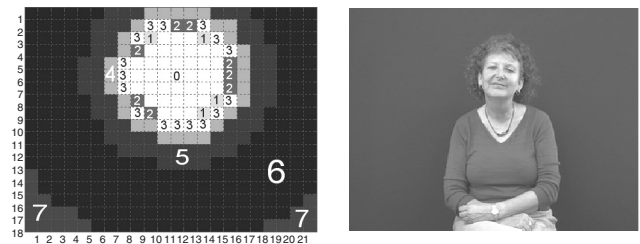


Fig. 2 Foveation MB map for one frame of clip 'FDD' with $v=3$, region numbers also shown

H.264 coding: The number of bits required for coding each MB in a video frame depends on the level of activity in the MB, the effectiveness of the prediction, and the quantisation step size. The latter is controlled by the quantisation parameter (QP). The value of QP controls the amount of compression and corresponding fidelity reduction for each MB. The foveation map described in the previous step is combined with a given range of QP values to produce MB regions that will be quantised with different step sizes, with the step size getting bigger for regions of higher eccentricity. More specifically, an algorithm was written which ensures that outer regions always have their QP increased before inner regions, and that the highest QP in the range is assigned to the lowest priority region. For example, for QP_{min} (minimum) = 30 and QP_{max} (maximum) ranging from 30 to 40 (i.e. for a QP range of 0 to 10) we get the QP allocations shown in Table 1 for the eight different foveation regions (region 0 is the highest priority region around the face the radius of which should be specified).

Table 1: Region QP assignments for different QP ranges

QP range	Region							
	0	1	2	3	4	5	6	7
0	30	30	30	30	30	30	30	30
1	30	30	30	30	30	30	30	31
2	30	30	30	30	30	30	31	32
3	30	30	30	30	30	31	32	33
4	30	30	30	30	31	32	33	34
5	30	30	30	31	32	33	34	35
6	30	30	31	32	33	34	35	36
7	30	31	32	33	34	35	36	37
8	30	31	32	33	34	35	36	38
9	30	31	32	33	34	35	37	39
10	30	31	32	33	34	36	38	40

When a variable QP (VQP) is used a small overhead is introduced in the bit stream due to coding of the different QP values of MBs lying on region borders. The overhead for the coded 'FDD' sequence (268 frames) with $v=3$ and QP range 30–40 was approximately 2.925 Kbits/s.

Table 2: Results of proposed VQP approach with and without (foveated) pre-filtering against results with constant QP (CQP), with same QP assignment for MBs corresponding to face region

'FDD'—foveated filtering (FV) + variable QP (VQP) (30–40)				
V	VQP rate (Kbit/s)	VQP reduction (%)	FV + VQP rate (Kbit/s)	FV + VQP reduction (%)
Original (CQP 30)	112.84	—	—	—
4	78.51	30.43	74.76	33.75
3	74.94	33.58	71.84	36.33
'Moving'—foveated filtering (FV) + variable QP (VQP) (30–40)				
V	VQP rate (Kbit/s)	VQP reduction (%)	FV + VQP rate (Kbit/s)	FV + VQP reduction (%)
Original (CQP 30)	211.51	—	—	—
4	137.52	34.98	125.15	40.83
3	130.67	38.22	118.17	44.13

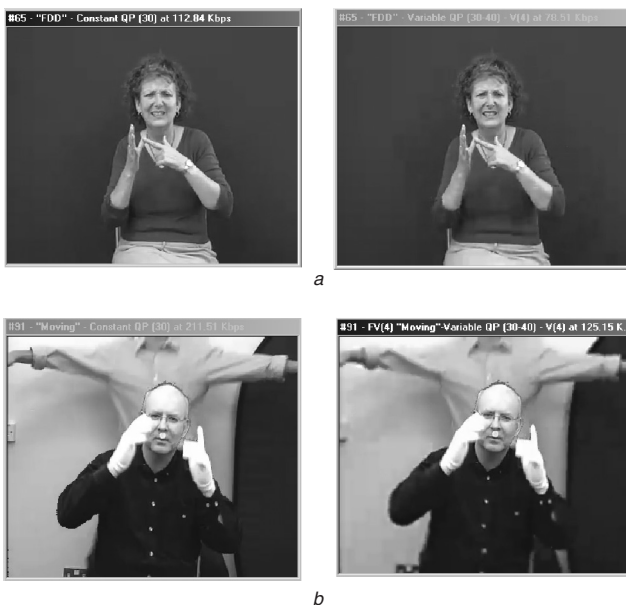


Fig. 3 One decoded frame coded with (left) constant QP = 30 (CQP) and (right) variable QP = 30–40 with $v = 4$ (VQP)

a 'FDD clip', VQP bit rate 30% less than CQP

b 'Moving clip', VQP + pre-filtering bit rate 40% less than CQP

Results: The H.264 reference software was modified to enable variable quantisation based on a given foveation map. The foveated processing unit which supplies the foveation map can also apply pre-filtering based on the same map. Two sequences were used, namely 'FDD,' a blue (plain) background sequence, and 'Moving', a moving background sequence. The output bitstreams conform to the baseline profile. The input frame rate was 25 fps, and the output frame rate 12.5 fps. Results for the two sequences are shown in Table 2. One decoded frame from each clip is shown in Fig. 3. It can be seen that a significant reduction of bit rate is achieved while keeping the quality of the important regions high (the face and the surrounding MBs). The subjective quality of the whole frame is also very good. Pre-filtering can offer approximately an additional 6% improvement in compression efficiency.

Conclusion: A foveated approach to sign language video coding is presented for lowering the bit rate requirements of sign language video without affecting significantly subjective quality (especially from a deaf viewer's point of view). Results show that a reduction of over 30% can be achieved while keeping the quality of important regions high. The proposed approach is based on the results of our eye tracking study which showed that experienced sign language viewers concentrate on the face and especially the mouth region.

© IEE 2003

3 October 2003

Electronics Letters Online No: 20031140

DOI: 10.1049/el:20031140

D. Agrafiotis, N. Canagarajah and D.R. Bull (*Image Communications Group, Centre for Communications Research, University of Bristol, Woodland Road, BS8 1UB, United Kingdom*)

M. Dye (*Centre for Deaf Studies, University of Bristol, Bristol BS8 2TN, UK*)

M. Dye: Now with Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA.

References

- 1 AGRAFIOTIS, D., *et al.*: 'Optimised sign language video coding based on eye-tracking analysis'. SPIE Int. Conf. on Visual Communications and Image Processing (VCIP), Lugano, Switzerland, July 2003
- 2 GEISLER, W.S., and PERRY, J.S.: 'A real-time foveated multiresolution system for low-bandwidth video communication', *SPIE Proc.*, 1998, **3299**
- 3 SHEIKH, H.R., *et al.*: 'Real time foveation techniques for H.263 video encoding in software'. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2001, Vol. 3, pp. 1781–1784