Complexity constrained rate-distortion optimization of sign language video using an objective intelligibility metric

Frank M. Ciaramello and Sheila S. Hemami

School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, 14853

ABSTRACT

Sign language users are eager for the freedom and convenience of video communication over cellular devices. Compression of sign language video in this setting offers unique challenges. The low bitrates available make encoding decisions extremely important, while the power constraints of the device limit the encoder complexity. The ultimate goal is to maximize the intelligibility of the conversation given the rate-constrained cellular channel and power constrained encoding device. This paper uses an objective measure of intelligibility, based on subjective testing with members of the Deaf community, for rate-distortion optimization of sign language video within the H.264 framework. The intelligibility metric is used in a Lagrangian cost function along with the Viterbi algorithm to make optimal mode and quantizer decisions for each macroblock. The optimal QP values are analyzed and the unique structure of sign language is exploited in order to reduce complexity by three orders of magnitude.

1. INTRODUCTION AND PREVIOUS WORK

Wireless and cellular video communication can offer the Deaf community the freedom of long distance communication in their native sign language (SL).[1] One key challenge is that the bandwidth available on cellular networks is very limited. Most traditional encoding techniques are optimized in terms of some measure of aesthetic distortion. As a communication tool, SL video must be judged in terms of its intelligibility; the desired outcome is that a viewer comprehends the linguistic information. Therefore, it is necessary to develop encoding algorithms that can maintain the intelligibility of SL communication while maximally compressing the video sequence to meet the stringent rate constraints. In addition to bandwidth constraints, the processing power is very limited for the majority of cellular devices. If the compression algorithm were too complex, the encoding process could not occur in real-time. The overall goal of the MobileASL project is to provide real-time intelligible SL communication over cellular telephones.[2] The encoding algorithms generate H.264 compliant bitstreams, to allow for the use of existing hardware decoders.

There is a significant amount of structure in SL that can be exploited in a video compression setting. All of the information in SL is conveyed through facial expressions and hand gestures.[3] Furthermore, eye tracking studies have revealed that when observing a signer, a fluent SL user will primarily gaze at the signer's face.[4–6] Because of this structure, SL video should be coded with high visual fidelity around the face, sufficiently high temporal resolution for capturing signs, and without many bits being spent on non-face, non-hand regions.[1,6]

There are many specialized algorithms for encoding SL video. Several methods involve transforming the sequence into a binary representation or line drawing.[7–9] While these algorithms achieve significant compression gains, they result in very unnatural videos. The goal of this work is to allow for signers to communicate naturally. Other algorithms exploit the inherent structure of SL for compression gains on natural video sequences. For example, Agrafiotis et al. use foveated processing to generate a map of priority regions.[5] By assigning macroblocks to each priority region, blocks nearest to the face are coded with more bits than blocks farther away. They achieved bitrate reductions of around 40% over the H.264 reference encoder (JM) without affecting the intelligibility of the sequence. Nakazono et al. proposed three techniques for improving SL compression in H.263: weighted bit allocation, modified macroblock encoding order, and forced SKIP mode in background blocks.[10] These techniques allowed more bits and higher priority given to the face and hand blocks. They demonstrated that at a fixed bitrate, their proposed algorithm had higher mean opinion scores (as rated by fluent SL users) than the H.263 test model. Other algorithms operate on the region-of-interest based approach.[11,12] These algorithms either arbitrarily assign more bits to the face and hand blocks or severely compress all non-skin blocks.

While all of these approaches exploit the inherent structure in SL videos, none are optimized with an appropriate cost function. The bit allocation to face and hand regions is done in a very heuristic fashion. The goal of this work is to implement a measure of intelligibility in a rate-distortion optimization setting. Analysis of the optimal coding parameters provides a method for substantially reducing complexity while achieving nearly optimal rate-distortion performance. Section 2 describes how the Viterbi algorithm is used with a Lagrangian cost function for rate-distortion optimization. Section 3 discusses the gains achieved by the optimized algorithm and how heuristics are applied to significantly reduce the complexity.

Presented at IEEE Western New York Image Processing Workshop (WNYIP), September 2007

2. ESTABLISHING OPTIMAL CODING PARAMETERS USING AN INTELLIGIBILITY METRIC

In H.264, the rate spent on a macroblock is determined by the selection of motion vector, mode, and quantizer.[13] The problem of rate control becomes choosing a parameter combination $p_i \in P \equiv \{MV \times M \times QP\}$ for each macroblock X_i over all N blocks. These coding decisions will affect total rate, $R(X,p)$, and total distortion, $D(X,p)$. Given a rate constraint, R_{max}, the optimization finds p such that:

$$\min_{p \in P^N} D(X,p) \quad \text{subj. to } R(X,p) \leq R_{max} \tag{1}$$

This rate-constrained optimization problem is made into an unconstrained problem by using the Lagrangian relaxation technique. The distortion metric used is based on a measure of intelligibility. In an earlier paper,[14] the authors developed a spatial distortion metric that correlates well with subjective intelligibility evaluation: $I = W_F MSE_F + W_H MSE_H$

This intelligibility metric is a weighted sum of MSE in the face pixels and MSE in the hand pixels. The optimal values for the weights are $W_F = 0.6$ and $W_H = 0.4$. Face and hand pixels are found using skin-color detection and morphological processing. In order to use this metric in an encoding setting, each macroblock in a frame is classified as either face, hand, or background according the pixel-level map. As measured, intelligibility is not affected by distortion in the background. However, ignoring background block distortion would reduce the optimization of those blocks to finding the parameters which result in the smallest rate. This would effectively encode all background blocks as skip type blocks and would result in very distracting artifacts. To appropriately handle this, it was experimentally verified that weighting background distortions by 10^{-4} sufficiently reduced the rate allocated to those regions while preventing extremely distorted compression artifacts.

As in the work of Wiegand et al.,[13] the following simplifications are made. First, for INTER mode macroblocks, the motion vectors are optimized, in the rate-distortion sense, before mode and QP decisions are made. Second, the mode decision can be optimized in the r-d sense, for a given QP.

The goal now is to find the optimal QP values for each macroblock in the frame, according to the Lagrangian cost. In H.264, the QP for the current block is coded as a delta offset from the QP for the previous block. Because of this, the additional rate required to encode large changes in QP can add significant overhead to the bitstream, especially at very low rates. In order to model this dependency, a trellis is built in which each stage corresponds to a macroblock and each node in a stage corresponds to a QP value.[15,16] The Viterbi algorithm is used to search for the path through the trellis that minimizes the Lagrangian cost.

This algorithm was implemented as a modified version of the x264 open source codec. This implementation of H.264 was selected because it consistently outperforms many other codecs in terms of both computation time and rate-distortion performance.[17]

3. R-D OPTIMIZED COMPRESSION AND COMPLEXITY REDUCTION

The optimization algorithm described in section 2 was applied to four different videos, three indoor scenes with a plain background and one outdoor scene with a very active background. The sequences 'Sakura' and 'Graduation' were recorded at University of Washington as part of the MobileASL project.[2] Both of these sequences were recorded at 15 fps and 320x240 pixels. 'Siblings' was taken from the American Sign Language Linguistic Research Project (ASLLRP) at Boston University[18] and is 320x240 pixels at 30 fps. 'Outdoor' was recorded at University of Bristol[5] and is 352x288 pixels at 25 fps.

Figure 1 illustrates the results for the sequence 'Outdoor' with the optimized algorithm, complexity-reduced algorithm, and with the standard x264 algorithm. In these plots, intelligibility is measured as $10 \log \frac{255^2}{I}$, which is the intelligibility distortion converted to quality score on a log scale.[14] In all four sequences, there is some gain in measured intelligibility over the standard rate control algorithm.

The most significant gains are obtained in the sequence 'Outdoor'. Because of the background activity, there is a large amount of residual left in background macroblocks after motion compensation. Since the x264 rate control algorithm chooses a single QP for the entire frame, it is forced to allocate a lot of bits to the background blocks, as well as the face and hand blocks. The SL optimized technique only assigns low QP values (and therefore more bits) to the face and hand macroblocks. This leads to improvements of 1-3 dB in intelligibility. The gains for the indoor sequences are not as drastic, because there is very little energy in the background macroblocks. The default encoder is already allocating almost all of the rate to the face and hand blocks, because there is almost no residual in the background after motion compensation. This leads to average gains of around 0.8 dB.

In terms of number of required Lagrangian cost calculations, the Viterbi algorithm has a time complexity of $O(52^2 \times M \times N)$, where there are M possible encoding modes, 52 possible QP values, and N macroblocks in a

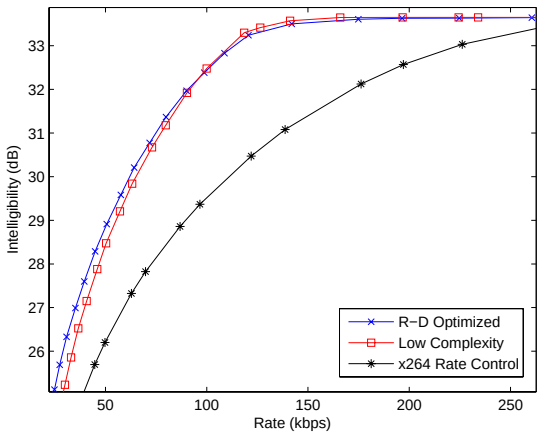
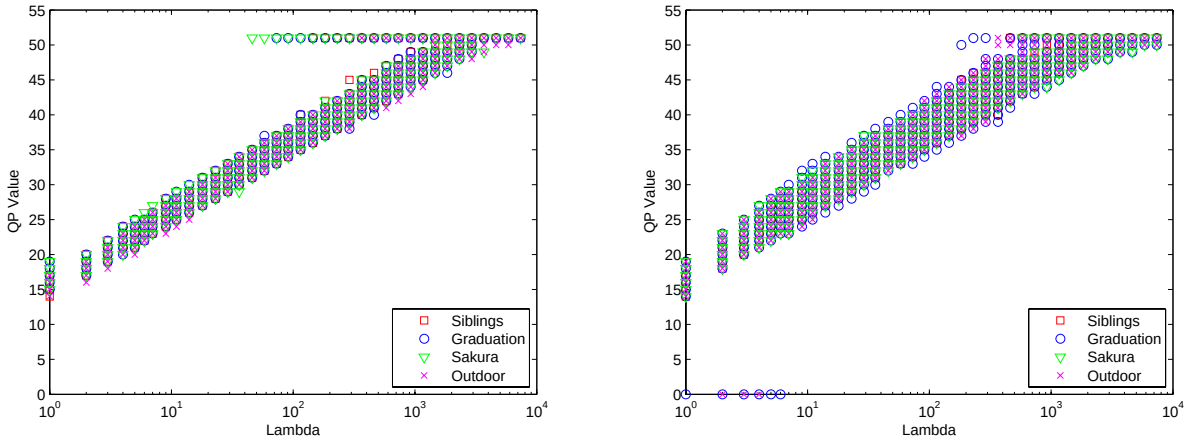


Figure 1. Plot of Intelligibility (dB) versus bitrate (kbps) for sequence 'Outdoor'. Optimal gains are as high as 2dB because the background is very active. Low complexity algorithm performs similar to the optimal case. The x264 rate control algorithm assigns a single QP to the entire frame, effectively adding unnecessary rate to background blocks.



(a) QP occurrence versus λ in non-SKIP mode face blocks. (b) QP occurrence versus λ in non-SKIP mode hand blocks.

Figure 2. A plot of optimal QP values selected versus lambda. 90% of the blocks had QPs among these clouds of points. The outliers in both (a) and (b) are a result of motion compensation.

frame. Because of this computational requirement, it is not feasible to use the Viterbi algorithm in a real-time scenario. However, the optimal QP selection can be analyzed to identify a relationship between λ and QP. Figure 2 illustrates the optimal QPs selected for a fixed value of λ in the face and hand blocks. 90% of the non-SKIP mode macroblocks have QPs among the plotted points. Only non-SKIP mode blocks are considered because QP has no meaning in the SKIP mode.

As seen in the figure, the range of optimal QPs selected is virtually identical in each sequence for a fixed λ, despite differences in framerate, resolution, and number of face and hand blocks. The outliers in both 2(a) and 2(b) are a result of motion compensation. In the face blocks, there is very little residual information left after motion compensation. On the other hand, there can be a significant amount of residual left in the hand blocks. Because of this, QP are adjusted higher and lower, respectively. By exploiting this relationship, the complexity-reduced encoding process simply uses a lookup table to select QP for a given λ and macroblock type (e.g. face, hand, or background). This is similar to the technique used by Wiegand and Girod[19] in which a functional relationship was developed so that λ could be calculated given a quantizer for the frame. Using a lookup table to select QP reduces the encoding complexity from $O(52^2 \times M \times N)$ to $O(M \times N)$, since only the lowest cost mode needs to be found at each macroblock.

The modified, low-complexity algorithm was implemented in x264 using the results from figure 2 by generating a QP lookup table. Given a specified value of λ, a QP is assigned to each macroblock type (face, hand, background) and that QP is applied to all blocks of that particular type. The results for the sequence 'Outdoor' are illustrated in figure 1. The low complexity algorithm performs nearly as well as the optimized method,

especially at higher rates. At the very low rates, the gains achieved by the low complexity method decrease compared to the R-D optimized method. As the total bitrate decreases, the rate added by coding large QP offsets has a larger impact. The dynamic programming formulation takes this into account by comparing all possible QP changes, but the low complexity method is forced to always choose the same QP, given a fixed λ.

Having established the relationship between λ and optimal QP values, the rate for a SL sequence can be controlled simply by adjusting λ and using the appropriate lookup table. Common rate control techniques involve measuring differences between frames and variance within blocks to estimate frame complexity and select an appropriate quantizer.[20] The SL compression setting allows for this method to be greatly simplified. Because all the important blocks will essentially contain the same type of content, part of a face or hand, the relationship between λ and the resulting bitrate will only depend on both the number of face and hand blocks and the framerate of the sequence. By properly modeling this dependency, rate-control can be employed efficiently in the low-complexity environment required by cellular devices.

REFERENCES

1. ITU-T, "Application profile - sign language and lip-reading real-time conversation using low bit-rate video communication," May 1999.
2. E. Riskin, S. Hemami, and R. Ladner, *The MobileASL Project*, http://www.cs.washington.edu/research/MobileASL/index.html.
3. S. K. Liddell and R. E. Johnson, "American sign language: The phonological base," in *Sign Language Studies*, **64**, pp. 195–278, 1989.
4. A. Cavender, R. Ladner, and E. Riskin, "Mobileasl: Intelligibility of sign language video as constrained by mobile phone technology," in *ASSETS 2006: The Sixth International ACM SIGACCESS Conference on Computers and Accessibility*, 2006.
5. D. Agrafiotis, N. Canagarajah, D. R. Bull, J. Kyle, H. Seers, and M. Dye, "A perceptually optimised video coding system for sign language communication at low bit rates," in *Signal Processing: Image Communication*, (21), pp. 531–549, 2006.
6. L. Muir, I. Richardson, and S. Leaper, "Gaze tracking and its application to video coding for sign language," in *Picture Coding Symposium 2003*, pp. 321–325, April 2003.
7. G. Sperling, M. Landy, Y. Cohen, and M. Pavel, "Intelligible encoding of asl image sequences at extremely low information rates," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, **24**, pp. 1061–1074, August 2002.
8. P. Letellier, M. Nadler, and J.-F. Abramatic, "The telesign project," in *Proceedings of the IEEE*, **73**, pp. 813–827, April 1985.
9. M. D. Manoranjan and J. A. Robinson, "Practical low-cost visual communication using binary images for deaf sign language," in *IEEE Trans. Rehabilitation Engineering*, **8**, pp. 81–88, March 2000.
10. K. Nakazono, Y. Nagashima, and A. Ichikawa, "Digital encoding applied to sign language video," in *IEICE Trans. Inf. & Sys.*, **E89-D**, June 2006.
11. D. M. Saxe and R. A. Foulds, "Robust region of interest coding for improved sign language telecommunication," in *IEEE Trans. Information Technology in Biomedicine*, **6**, pp. 310–316, December 2002.
12. R. Schumeyer, E. Heredia, and K. Barner, "Region of interest priority coding for sign language videoconferencing," in *IEEE Multimedia Signal Processing Workshop*, pp. 531–536, June 1997.
13. T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," in *IEEE Trans. Circuits and Systems for Video Technology*, **13**, July 2003.
14. F. Ciaramello and S. Hemami, "'Can you see me now?' An objective metric for predicting intelligibility of compressed American Sign Language video," in *Proc. SPIE Vol. 6492, Human Vision and Electronic Imaging '07*, B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds., **6492**, 2007.
15. A. Ortega and K. Ramchandran, "Forward-adaptive quantization with optimal overhead cost for image and video coding with applications to mpeg video coders," in *Proc. of IS&T/SPIE Digital Video Compression '95*, February 1995.
16. G. M. Schuster and A. K. Katsaggelos, "Fast and efficient mode and quantizer selection in the rate distortion sense for H.263," in *Proc. SPIE Vol. 2727, p. 784-795, Visual Communications and Image Processing '96, Rashid Ansari; Mark J. Smith; Eds.*, R. Ansari and M. J. Smith, eds., **2727**, pp. 784–795, Feb. 1996.
17. D. Vatolin, D. Kulikov, A. Parshin, A. Titarenko, and M. Smirnov, *Moscow State University MPEG-4 AVC/H.264 Video Codec Comparison*, http://compression.ru/video/codec_comparison/mpeg-4_avc_h264_2006_en.html, November 2006.
18. C. Neidle and S. Sclaroff, *American Sign Language Linguistic Research Project [Online]*, http://ling.bu.edu/asllrpdata/queryPages/.
19. T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control," in *Proceedings of International Conference on Image Processing*, **3**, October 2001.
20. I. E. G. Richardson, *Video Codec Design*, John Wiley & Sons Ltd, West Sussex PO19 IUID, England, 2002.