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. Performance bounds are established by using the intelligibility metric in a Lagrangian cost function along with a trellis search 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 relative to the trellis search technique with no loss in rate-distortion performance. Further reductions in complexity are made by eliminating rarely occuring modes in the encoding process. The low complexity SL optimization technique increases the measured intelligibility up to 3.5 dB, at fixed rates, and reduces rate by as much as 60% at fixed levels of intelligibility with respect to a rate control algorithm designed for aesthetic distortion as measured by MSE.

Keywords: Sign language video coding, H.264, complexity constrained video coding, rate-distortion optimization

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, typically MSE. 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, encoding algorithms are necessary 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 is too complex, the encoding process cannot 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 developed generate H.264 compliant bitstreams, to allow for the use of existing hardware decoders.

Sign language contains a significant amount of structure that can be exploited in a video compression setting. All of the information in SL is conveyed through facial expressions and hand gestures.[3,4] Furthermore, several eye tracking studies have revealed that when observing a signer, a fluent SL user will primarily gaze at the signer's face.[5–7] This phenomenon occurs because subtle changes in facial expression can substantially change the meaning of a hand gesture.[8,9] A gaze in a particular direction can indicate a pronoun and raising one's eyebrows indicates a question. 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.[1,7]

Many specialized algorithms have been proposed for encoding SL video. Several methods involve transforming the sequence into a binary representation or line drawing. Both the Telesign Project[10] and Sperling et al.[11]

F.M.C: E-mail: fmc3@cornell.edu; S.S.H: E-mail: hemami@ece.cornell.edu

converted sign language into cartoon-style line drawings. The Telesign Project captured single signs at two different video parameters: 256×256 at 12.5 fps and 128×128 at 25 fps. The signs were encoded at 75.2 kbps and 38.6 kbps (corresponding to the two sets of video parameters) and subjective observers correctly identified over 90% of the signs. Sperling et al. encoded 15 fps sequences at rates between 3 and 15 kbps. Subjective observers identified 70-80% of the individual signs. More recently, Manoranjan and Robinson used binary sketches at four resolutions (320×240, 120×160, 160×120, and 80×60) to transmit sign language conversations at a fixed rate of 33.6 kbps.[12] Subjective evaluation found that users preferred the smallest resolution, primarily because these were encoded at the highest framerate (7-8 fps). While these algorithms maintain intelligibility at low bitrates, they result in very unnatural videos. Furthermore, they require encoding and decoding techniques that are specialized for binary video sequences, making implementation on complexity-constrained devices difficult.

More recent SL compression algorithms (including this work) use specialized processing while still conforming to pre-existing encoding standards such as H.26x and MPEG. These algorithms exploit the inherent structure of SL for compression gains on natural video sequences and also allow the data streams to be decoded on widely available, standards-compliant decoders. One common approach to SL encoding is to apply region-of-interest based compression. Both Schumeyer et al.[13] and Saxe et al.[14] use automatic skin segmentation techniques to identify the region-of-interest. These algorithms assign more bits to the face and hand blocks by adjusting quantizer values and severely compressing all non-skin blocks. Schumeyer et al. encoded QCIF (176×144) sized images at fixed bitrates of 64 kbps and 128 kbps using H.261. Reductions of 10-15% in the number of bits per picture led to slight increases in the effective framerate, relative to an encoding technique that assigns a uniform quantizer to the entire frame. The effective framerates for the sequences were 16.3 fps at 64 kbps and 17.7 fps at 128 kbps. Saxe et al. add an additional preprocessing step to blur all the non-skin regions, with the intention of reducing blockiness in the background regions. MPEG-1, motion JPEG, and the Windows Media Encoder were applied to video sequences recorded at 30 fps with a resolution of 160×120. By reducing the quality in the background region, bitrates were reduced by 25% over the cases in which no region-of-interest coding was used. In both cases, no formal study was performed to verify intelligibility.

Nakazono et al. proposed three techniques for improving SL compression in H.263: weighted bit allocation, modified macroblock processing order, and forced SKIP mode in background blocks.[15] The weighted bit allocation decreases the rate allocated to each macroblock as a function of increasing distance from the face. The modified processing order adjusts the analysis of blocks, such that blocks near the face are analysed first. The encoder will obtain information about the face blocks earlier in the encoding process. Finally, a set of background macroblocks at the edges of the frame are identified and are always encoded in the SKIP mode. These techniques allowed more bits to be assigned to the face and regions near the face, but requires that the weights and block labeling are manually tuned prior to encoding. The source content used was 15 fps sequences at both CIF (352×288) and QCIF resolutions. They demonstrated that at fixed bitrates of 256 kbps, 128 kbps, and 64 kbps, their proposed algorithm had higher mean opinion scores (as rated by fluent SL users) than the H.263 test model.

Agrafiotis et al. use foveated processing to generate a map of priority regions.[6] The face is identified automatically using skin segmentation and facial feature detection. Given the location of the face, a foveation model is used to assign macroblocks to each priority region. A different quantization parameter is assigned to each priority region, allowing blocks nearest to the face to be coded with more bits than blocks farther away. These modifications conform to the H.264 standard and were applied to four CIF size sequences recorded at 25 fps. At average rates of 132 kbps, they achieved an average bitrate reduction of 40% over the H.264 reference encoder (JM) without affecting the intelligibility of the sequence. This technique appropriately considers how sign language is viewed and heuristically determines how much rate should be given to each of the priority regions.

While all of these approaches exploit the inherent structure in SL videos, none are optimized with a cost function that measures intelligibility. The additional rate allocated to the face and hand regions is selected heuristically. Furthermore, a non-trivial amount of rate is allocated to the background region, which is not required for intelligible sign language. The goal of this work is to implement an objective measure of intelligibility in a rate-distortion optimization setting, which provides a performance upper bound that inherently exploits the structure of SL. A low complexity algorithm, developed from the analysis of the resulting coding parameters,

achieves the same rate-distortion performance. Section 2 describes how a trellis-based model is used with a Lagrangian cost function for rate-distortion optimization. The cost function formulation includes a distortion measure that correlates well with subjective evaluation of intelligibility. Section 3 discusses the gains achieved by the optimized algorithm and how heuristics are applied to significantly reduce the complexity. Finally, the low complexity implementation is compared with the foveation-based SL optimization.[6]

2. ESTABLISHING PERFORMANCE BOUNDS AND 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.[16] 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. This reduces the optimization in Equation (1) to:

$$\min_{p \in P^N} J(X,p) = D(X,p) + \lambda R(X,p) \tag{2}$$

The distortion metric used here is based on a measure of intelligibility. In an earlier paper,[17] 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 values for the weights that maximize correlation with subjective intelligibility ratings are $W_F = 0.6$ and $W_H = 0.4$. These values are supported by the linguistic analyses of sign language[8,9] and by the eye tracking studies.[5–7] Subtle details in the face are essential for accurate intelligibility of a sign language conversation.

Face and hand pixels are found on each frame 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, the authors verified that weighting background distortions by 10^{-2} sufficiently reduced the rate allocated to those regions while preventing extremely distorted compression artifacts.

As in the works of Wiegand et al.,[18] the selection of the parameters p is further simplified. 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 rate-distortion sense, for a given QP.

The goal then becomes 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 in a row and each node in a stage corresponds to a QP value.[19,20] The Viterbi algorithm is used to search for the path through the trellis that minimizes the Lagrangian cost for a particular row. The algorithm then iterates over all rows in the frame. In terms of number of required Lagrangian cost calculations, this algorithm has a 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 frame.

In addition to the dependencies within a row, many of the prediction modes in H.264 lead to depencies across rows. However, the computational complexity required to fully model these dependencies is $O(52^4 \times M \times N)$, making it impossible to evaluate in any practical amount of time. The results of the trellis search algorithm will

Table 1. Summary of resolution, framerate, and region details for each sign language sequence, with resolutions reported in terms of macroblocks. The face, hand, and background sizes are all per frame averages and the numbers in parenthesis correspond to the percentage of the frame corresponding to that region. There are more face blocks in the sequence 'Outdoor' because the signer is closer to the camera in this case. The sequences 'Siblings' and 'Outdoor' have fewer hand blocks than face blocks because only the upper body of the signer was filmed.

Sequence	Macroblocks	Framerate	Face Size	Hand Size	BG Size	BG Activity
Sakura	20x15 (320×240)	15 fps	13.5 (4%)	18.6 (6%)	267.8 (90%)	Static
Graduation	20x15 (320×240)	15 fps	13.6 (4%)	18.9 (6%)	267.5 (90%)	Static
Siblings	20x15 (320×240)	30 fps	17.2 (6%)	6.7 (2%)	276.0 (92%)	Static
Outdoor	22x18 (352×288)	25 fps	38.9 (10%)	16.9 (4%)	340.2 (86%)	High

be considered the computable upper bound for rate-distortion performance. Distortion in this case is measured by I, as described above. This algorithm was implemented as a modified version of the x264 open source codec. This particular implementation of H.264 was selected because it consistently outperforms many other codecs.* The relevant portions of the rate control employed by x264 are highlighted in the Appendix. A detailed description can be found in the work by Merritt and Vanam.[23]

3. OPTIMAL CODING DECISIONS AND COMPLEXITY REDUCTION

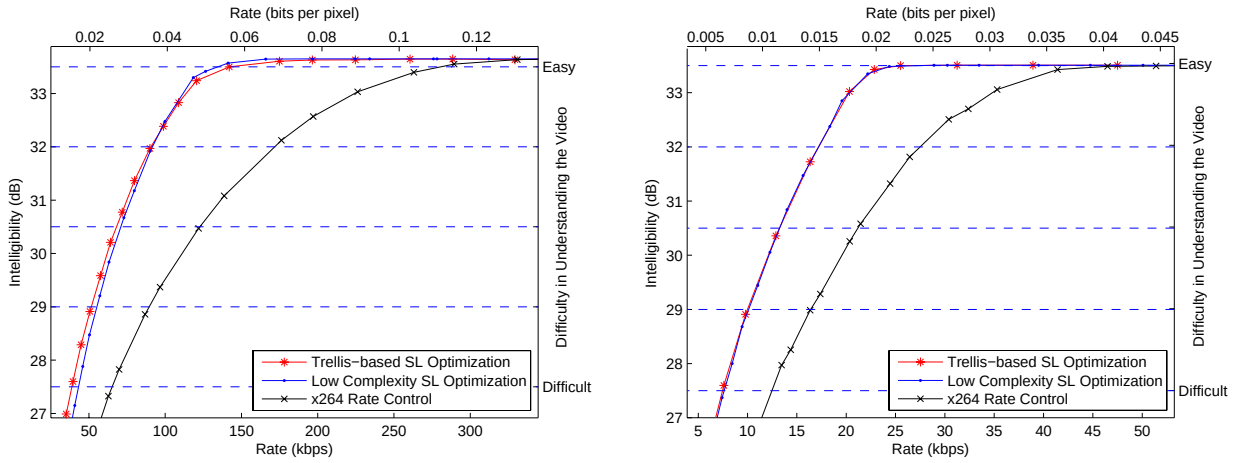
The optimization algorithm described in Section 2 was applied to four different videos, three indoor scenes with a plain background and one outdoor scene with a very active background. The sequences 'Sakura' and 'Graduation' were recorded at University of Washington as part of the MobileASL project.[2] 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[24] and is 320x240 pixels at 30 fps. 'Outdoor' was recorded at University of Bristol[6] and is 352x288 pixels at 25 fps. Table 1 contains a summary of these properties.

The trellis-based optimization algorithm sets a computable upper bound on rate-distortion performance for the sign language sequences. The results for intelligibility are reported as $10 \log \frac{255^2}{I}$, which is the intelligibility distortion, I, converted to quality score on a log scale.[17] In all four sequences, there is some gain in measured intelligibility over the x264 rate control algorithm. Table 1 illustrates several interesting differences in the content of each sequence that have a direct impact on the compression gains achieved. In the three indoor sequences, the signer is further from the camera than in the outdoor sequence, which is evident in the higher percentage of face macroblocks in 'Outdoor.' As a result, the sequence 'Outdoor' requires a higher rate for the same level of intelligibility, as seen in Figure 1.

Also note that both 'Sakura' and 'Graduation' have more hand blocks than face blocks. The signer in these sequences is wearing a short sleeve shirt. The face and hand detection algorithm classifies the entire arm as part of the hand and which results in more hand blocks. Furthermore, in both sequences the signers' hands were always in the frame. In 'Siblings' and 'Outdoor,' only the signers' upper body was filmed, resulting in many frames where only a single hand was visible. Many signs in sign language only require the use of one hand; the secondary hand is often at rest and, in these cases, off camera.

In all of the sequences, the trellis-based optimization technique achieves a 2-3 dB increase in intelligibility at a fixed rate. The largest differences between the sequences are in the reduction of rate at a fixed level of intelligibility with respect to the x264 rate control method. The level of activity in the background has the most prominent effect on these compression gains. The indoor sequences all had similar compression gains of 20-50% over the x264 rate control; the amount of compression varied with intelligibility. The outdoor sequence had gains of 40-60%. Figure 1 illustrates the compression results for sequences 'Outdoor' and 'Sakura' with the trellis-based SL optimization, the low complexity SL optimization, and the standard x264 rate control algorithm.

*A study performed at Moscow State University compared several H.264 codecs when applied to various natural video scenes.[21] The codecs were analyzed in terms of both computation time and rate-distortion performance, where distortion in this case was measured as both PSNR and SSIM.[22]



(a) Rate (kbps) versus Intelligibility (dB) for sequence 'Outdoor'. Gains are between 2dB and 3.5dB.

(b) Rate (kbps) versus Intelligibility (dB) for sequence 'Sakura'. Gains are between 2dB and 3dB

Figure 1. Plots of Intelligibility (dB) versus bitrate (kbps and bpp). The secondary y-axis maps the measured intelligibility to the five-point subjective scale used in a previous work.[17] The five-point scale is in response to the question "How easy or how difficult was it to understand the video?" The gains in (a) are much larger because the background is very active. The x264 rate control algorithm assigns a single QP to the entire frame, effectively adding unnecessary rate to background blocks.

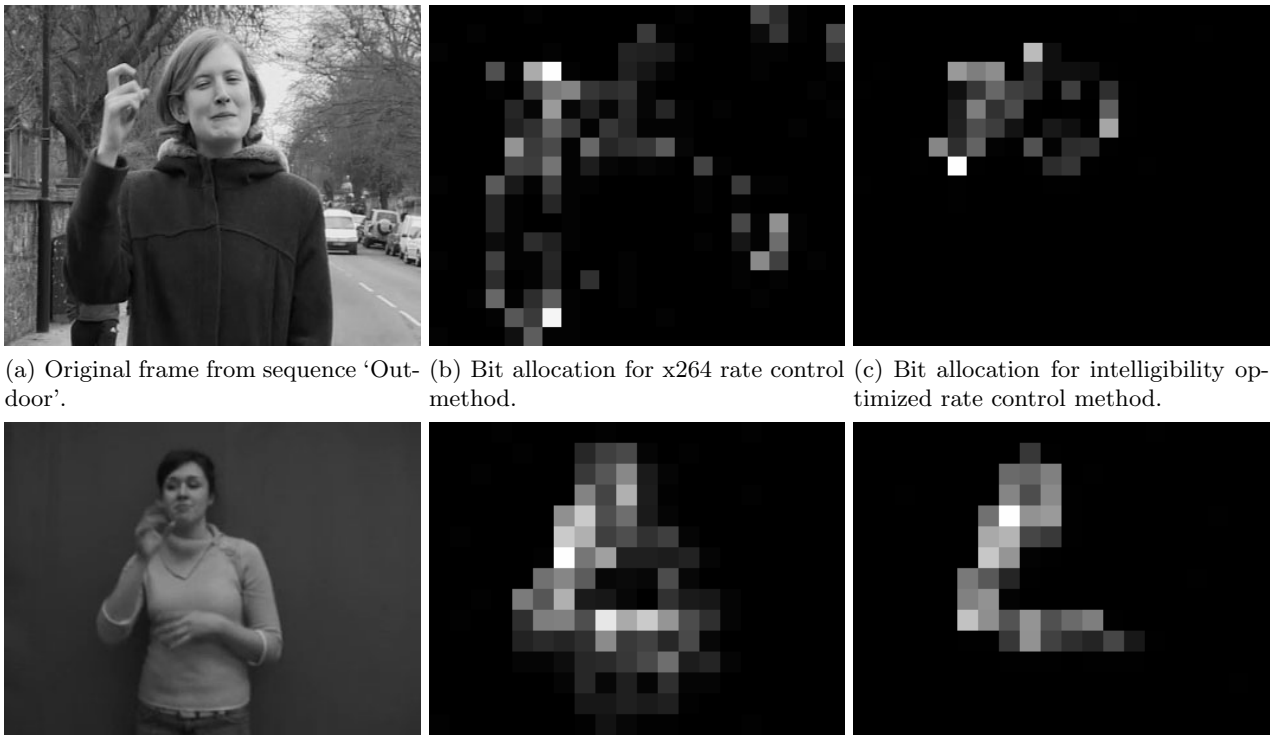
The dashed lines in the figure map the objective intelligibility measure I to the subjective scale used in the author's previous work.† When intelligibility becomes "easy", the proposed SL optimization method reduces the rate by over 50% versus the x264 rate control, as seen in Figure 1. Furthermore, any increase in rate beyond the point when intelligibility is easy does not increase an observer's ability to understand the SL content. For a rate-constrained communication channel (i.e., cellular networks), this implies that only enough rate needs to be allocated to each user such that their conversation is easy to understand. This can relieve the overall system of unnecessary encoding and transmission.

The most significant compression gains are obtained in the sequence 'Outdoor'. Because of the higher background activity, a large amount of residual energy remains in the background macroblocks after motion compensation. Since the x264 rate control algorithm chooses a single QP for the entire frame, it is forced to allocate bits to the background blocks, as well as the face and hand blocks. The SL optimized techniques (trellis-based and low complexity) only assign low QP values (and therefore more bits) to the face and hand macroblocks. The residual coefficients in the background blocks are severely quantized. The gains for the indoor sequences are slightly smaller. Because of the low levels of background activity, motion compensantion results in mostly very small transform coefficients in background macroblocks. Because there is very little residual energy, it takes very little rate to encode, even at low QP values. As a result, the default encoder is already allocating almost all of the rate to the face and hand blocks. Sample bit allocation results are demonstrated in Figure 2.

3.1. Low Complexity Heuristics

Ultimately, the sign language encoding algorithm will run in real-time on a complexity constrained device (e.g., a cellular phone). As mentioned in Section 2, using the trellis search to find the optimal QP values results in a time complexity of $O(52^2 \times M \times N)$. Because of this large computational requirement, it is not feasible to use the trellis-based algorithm in a real-time scenario. However, the optimal QP selections can be analyzed to identify a relationship between λ and QP. Figure 3 illustrates the optimal QPs selected for a fixed value of λ

†Participants viewed compressed sign language sequences and were asked the question "How easy or how difficult was it to understand the video?" They responded on a 5-point Likert scale between "Difficult" and "Easy."[17]



(a) Original frame from sequence 'Outdoor'. (b) Bit allocation for x264 rate control method. (c) Bit allocation for intelligibility optimized rate control method.

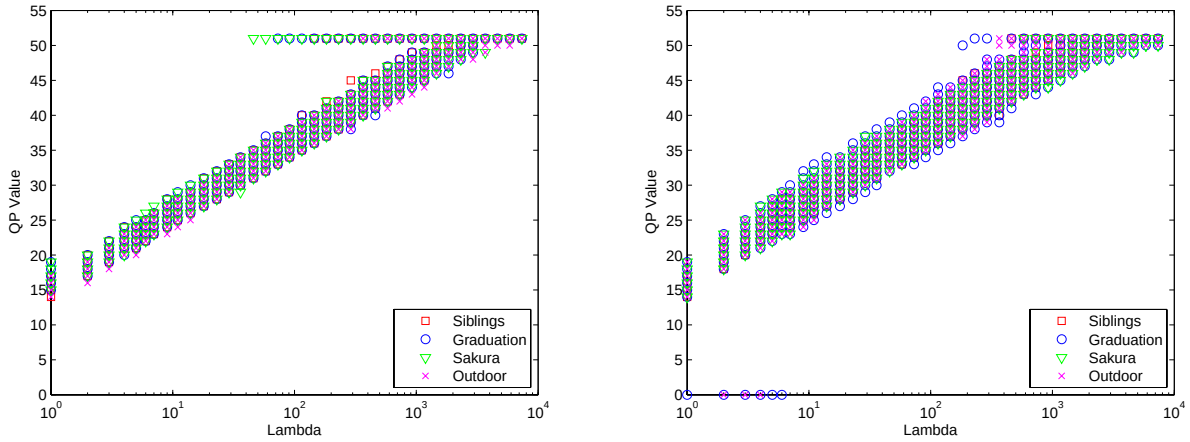
(d) Original frame from sequence 'Sakura'. (e) Bit allocation for x264 rate control method. (f) Bit allocation for intelligibility optimized rate control method.

Figure 2. Bit allocation for each rate control method. The brightness of a block corresponds to the amount of rate allocated to a macroblock relative to the entire frame. Notice that for the intelligibility optimized sequences (c and f), nearly all the rate is allocated to the face and hands. In the 'Outdoor' scene with an active background, the x264 rate control is allocating significant rate to regions that are unimportant to intelligibility, e.g. the trees and the person walking in the background.

in the face and hand blocks. Ninety percent 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 3(a) and 3(b) are a result of motion compensation. In the face blocks, because of the relatively low activity, very little residual information is left after motion compensation. On the other hand, a significant amount of residual energy remains 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 the QP for a given λ and macroblock type (e.g. face, hand, or background). The QP lookup table is generated from the average QP selected at each λ. The QP for background blocks was nearly always set to 51, the highest supported by H.264. This is similar to the technique used by Wiegand and Girod[25] 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 3 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 sequences 'Outdoor' and 'Sakura' are illustrated in Figure 1. In all the sequences, the low complexity algorithm performs as well as the trellis-based optimization method.



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

Figure 3. A plot of optimal QP values selected versus lambda. Ninety percent (90%) of the blocks had QPs among these clouds of points. The outliers in both (a) and (b) are a result of motion compensation. Background blocks are not included in the plot, as the trellis search chose a QP of 51 for nearly all background macroblocks. The average QP is used in the low complexity lookup table.

By using a lookup table for selecting the quantization parameter, the encoding complexity for the intelligibility optimized algorithm and the x264 default rate control method is equivalent. In H.264, as many as 12-15 different modes need to be analyzed for any given macroblock, which can add a significant amount of encoding time. The complexity of the encoding process is further reduced by analyzing the histogram of selected modes for each sequence and restricting the mode search to only those modes that are selected frequently. The restricted modes are established prior to encoding and, as a consequence, must be independent of a particular rate. The face and hand histograms revealed that each of the modes were selected at least 10% of the time for a subset of rates. For example, in the sequence 'Sakura,' the Intra 4x4 mode was selected on over 20% of hand blocks at rates above 18 kbps but only on 2% of the blocks at rates below 10 kbps. Because the number of face and hand blocks is only between 10% and 15% of the total number of macroblocks, it is acceptable to evaluate all the possible modes. Complexity reductions on the background blocks have a more significant effect on overall encoding complexity.

In both the high and low background activity cases, the finest Inter prediction modes were selected on fewer than 1% of the background blocks, regardless of rate. Removing these modes (Inter 4x4, 4x8, and 8x4) from the encoding process results in no change in rate-distortion performance. Of the remaining mode types, 90% of the background macroblocks were limited to Inter 16x16, Intra 16x16 partitions, and the SKIP mode. This implies that only the coarsest mode types need to be evaluated for many of the background blocks. Figure 4 illustrates the resulting rate-distortion performance for the mode restrictions for the sequences 'Sakura' and 'Outdoor'. The rates required for intelligibility levels of "easy" are summarized in Table 2

For the low framerate indoor sequences, restricting the mode search to only these three types led to a 6% rate increase for "easy to understand" sequences. The mode restricted algorithm requires a 14-16% increase in rate for the outdoor sequence and the 30 fps indoor sequence to be "easy to understand". The rate increase occurs because in many cases, allowing the use of Intra 4x4 and of the Sub-16x16 Inter prediction modes (16x8, 8x16, 8x8) resulted in a block with all zero-valued transform coefficients. H.264 has syntax specified for these cases and can encode them with very little rate. At higher framerates, the increased correlation between frames allows the motion compensation to find better block matches, which increases the occurrence of the zero-valued blocks. Thus, the higher framerate sequences are more affected by the mode restriction. While this mode restriction leads to the slight rate-distortion performance decrease, it also provides a significant reduction in complexity. By allowing only 16x16 partitions and the SKIP prediction mode in the background blocks, the complexity is reduced from $O(M \times N_{BG})$ to $O(3 \times N_{BG})$, where M is typically around 12 and N_{BG}, the number of background

Table 2. Summary of the rate at which intelligibility is "easy" for the low complexity SL optimization with and without mode restrictions, the foveation-based SL optimization, and the x264 rate control. The number in parenthesis is the relative reduction in rate from the x264 rate control. Note that for the sequence 'Siblings,' the foveation-based algorithm performs as well as the proposed low complexity algorithm. Because 'Siblings' has a small percentage of hand macroblocks and a very static background (Table 1), both the foveated technique and the proposed low complexity technique allocate most of the rate to the face region.

Sequence	Low Complexity	Mode Restricted	Foveation-based	x264 Rate Control
Sakura	27.5 kbps (46.5%)	29 kbps (43.6%)	48.2 kbps (6.2%)	51.4 kbps (0%)
Graduation	26.5 kbps (57.9%)	28.3 kbps (55.1%)	51.0 kbps (19.0%)	63.0 kbps (0%)
Siblings	41.5 kbps (46.1%)	48.8 kbps (36.6%)	42.0 kbps (45.5%)	77.0 kbps (0%)
Outdoor	118.0 kbps (57.9%)	134.0 kbps (52.1%)	252.0 kbps (10.0%)	280.0 kbps (0%)

blocks, is 90% of the frame.

3.2. Comparison with Foveation-based SL Optimization Algorithm

The x264 rate control algorithm is designed to minimize aesthetic distortion, measured in terms of MSE. To be fair, the proposed low complexity SL optimization technique is also compared against the foveation-based SL optimization proposed by Agrafiotis et al.[6] The foveation-based algorithm identifies the face and defines this as priority region 0. Seven more priority regions are defined based on a foveation model with a fixation point in the center of the face. A fixed QP is assigned to region 0 and this QP is incremented by two in each of the successive regions.

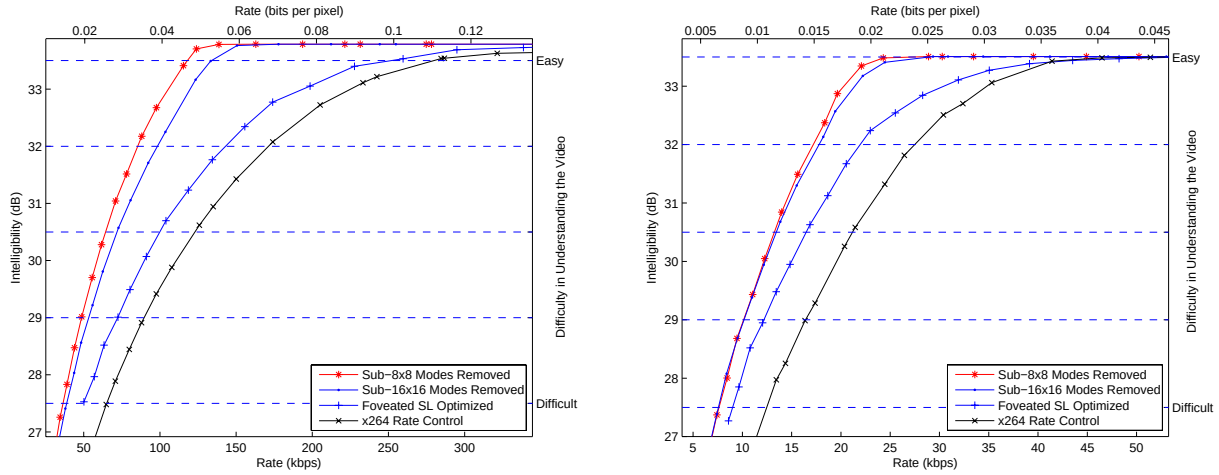
Figure 4 compares the performance of the foveation-based algorithm with the proposed low complexity algorithm. Because the foveation-based algorithm is allocating more rate to regions near the face, intelligibility is increased by about 1 dB (or about 0.67 points on the Likert scale) over the x264 rate control at a fixed rate. However, a large amount of rate is still allocated to background macroblocks, even though the information in these blocks adds nothing to the intelligibility of the sign language content. The proposed algorithm spends very little rate on background blocks, making more bits available for the face and hand regions. As a result, the proposed algorithm reduces the rate at which intelligibility is "easy" by an average of 38% with respect to the foveated-based approach. The results for each sequence are summarized in Table 2.

4. SUMMARY

A specialized encoding algorithm for sign language (SL) video was developed. The algorithm exploits the structure of SL by identifying the face and hand macroblocks and using an objective measure of intelligibility in a rate-distortion setting. A trellis search was used to identify optimal quantization parameters for each of the block types (face, hand, background). This computable upper bound leads to a method for selecting QP from a lookup table based on a given value of λ. Further complexity gains were achieved by limiting the mode selection process for background blocks. In terms of measured objective intelligibility, this algorithm achieves as high as 3.5 dB improvement over the x264 rate control technique used by x264 and 2 dB over a foveation-based SL optimization technique. Compression gains of over 50% versus the x264 rate control are achieved when the videos are easily intelligible.

5. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Dimitris Agrafiotis and Dr. David Bull for their willingness to share sign language sequences and for providing access to their algorithm. Their contributions improved the value of this work.



(a) Rate (kbps) versus Intelligibility (dB) for sequence 'Outdoor'.

(b) Rate (kbps) versus Intelligibility (dB) for sequence 'Sakura'.

Figure 4. Plots of Intelligibility (dB) versus bitrate (kbps and bpp). The secondary y-axis maps the measured intelligibility to the five-point subjective scale used in a previous work.[17] The five-point scale is in response to the question "How easy or how difficult was it to understand the video?" Restricting the background blocks to only three modes resulted in almost no loss in performance for the indoor sequence in (b). The foveated based SL optimization[6] gains about 1dB in intelligibility over the x264 rate control.

6. APPENDIX

The goal of the one-pass x264 rate control algorithm is to produce videos with consistent MSE across frames at a specified average bitrate. This method is mostly empirical. The relative amount of rate to be allocated to each frame (independent of the target bitrate) is proportional to a heuristic measure of complexity. Frames with high complexity will receive a higher percentage of rate while frames with low complexity receive less. Fast motion estimation is performed on a half-resolution copy of the frame and the complexity for the frame is calculated as the sum of absolute differences in the transformed residuals. The relative bits between the frames is empirically selected as $rbits = complexity^{0.6}$. This relative number of bits is then scaled to produce the total bits to be allocated to the frame. The scaling factor is calculated to be the value that would have resulted in the target bitrate, if it had been applied on all the previous frames. This effectively averages the complexity over all the previously encoded frames and chooses a scaling factor accordingly. Compensation for prediction errors is achieved by multiplying the bit allocation by $\frac{target\ filesize}{real\ filesize}$. Given this bit allocation, a single quantization parameter (QP) is then selected for the entire frame. Mode selection is performed using a Lagrangian cost function where distortion is measured as sum of squared differences (SSD).

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. W. C. Stokoe, *Sign Language Structure*, Linstok Press, Inc, Silver Spring, MD, 1978.
4. S. K. Liddell and R. E. Johnson, "American sign language: The phonological base," in *Sign Language Studies*, **64**, pp. 195–278, 1989.
5. 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.

6. 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.
7. 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.
8. C. Baker and C. A. Padden, "Focusing on the nonmanual components of American Sign Language," in *Understanding language through sign language research*, P. Siple, ed., pp. 27–57, Academic Press, (New York, NY), 1978.
9. S. K. Liddell, "Nonmanual signals and relative clauses in American Sign Language," in *Understanding language through sign language research*, P. Siple, ed., pp. 59–90, Academic Press, (New York, NY), 1978.
10. P. Letellier, M. Nadler, and J.-F. Abramatic, "The Telesign Project," in *Proceedings of the IEEE*, **73**, pp. 813–827, April 1985.
11. G. Sperling, M. Landy, Y. Cohen, and M. Pavel, "Intelligible encoding of ASL image sequences at extremely low information rates," in *Computer Vision, Graphics, and Image Processing*, **31**, pp. 335–391, 1985.
12. 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.
13. 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.
14. 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.
15. K. Nakazono, Y. Nagashima, and A. Ichikawa, "Digital encoding applied to sign language video," in *IEICE Trans. Inf. & Sys.*, **E89-D**, June 2006.
16. 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.
17. 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.
18. T. Wiegand, M. Lightsone, D. Mukherjee, T. G. Camplbell, and S. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard," in *IEEE Trans. Circuits and Systems for Video Technology*, **6**, April 1996.
19. 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.
20. 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.
21. 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.
22. Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," in *Signal Processing: Image Communication special issue on Objective video quality metrics*, **19**, pp. 121–132, February 2004.
23. L. Merritt and R. Vanam, "Improved rate control and motion estimation for H.264 encoder," *Image Processing, 2007. ICIP 2007. IEEE International Conference on* **5**, pp. V –309–V –312, Sept. 16 2007-Oct. 19 2007.
24. C. Neidle and S. Sclaroff, *American Sign Language Linguistic Research Project [Online]*, http://ling.bu.edu/asllrpdata/queryPages/.
25. T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control," in *Proceedings of International Conference on Image Processing*, **3**, October 2001.