THE INFLUENCE OF SPACE AND TIME VARYING DISTORTIONS ON OBJECTIVE INTELLIGIBILITY ESTIMATORS FOR REGION-OF-INTEREST VIDEO

Frank M. Ciaramello and Sheila S. Hemami

School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14853.
fmc3@cornell.edu, hemami@ece.cornell.edu

ABSTRACT

Objective estimators for video are expected to estimate accurately subjective ratings provided by humans. This work presents a subjective experiment designed to acquire intelligibility ratings for a collection of compressed ASL videos. The distortions present in the experimental database are analyzed in terms of their impact on the performance of objective estimators. Distortions that do not significantly vary across space or time cannot adequately challenge traditional objective estimators, such as PSNR and RMS distortion contrast, and an objective intelligibility measure designed specifically for ASL video provides negligible improvements in prediction accuracy. Distortions that vary across space and time, affecting only localized regions in the video, are considered spatially and temporally diverse. When the distortions present in the experimental database are sufficiently diverse, the objective intelligibility measure estimates subjective ratings more accurately than PSNR and RMS distortion contrast.

Index Terms— Region-of-interest coding, sign language video, video quality assessment, video quality database

1. INTRODUCTION

Real-time, two-way transmission of American Sign Language (ASL) video over cellular networks provides natural communication among members of the Deaf community. Region-of-interest (ROI) techniques suit ASL video, because the signer is the ROI. Subjective experiments provide an accurate evaluation of the quality of service being delivered by a such a video compression and transmission system. Because of the high cost of subjective studies, objective assessment techniques are developed to estimate the subjective ratings provided by humans and are evaluated in terms of their accuracy on a specific test database.

The experimental test database analyzed in this work is comprised of compressed ASL video, encoded at very low bitrates using region-of-interest encoders designed for ASL content. The subjective experiment that produces the test database is presented in Section 2. Four objective estimators are studied in terms of their ability to estimate subjective intelligibility scores. However, any conclusions regarding the performance of an objective estimator must be conditioned on the content available in the test database. In order to confidently compare objective estimators, the distortions contained in the experimental data must be evaluated in terms of their ability to adequately challenge the objective estimators. Section 3 analyzes the video content and distortions present in the experimental database and studies their impact on the performance of the objective estimators. Concluding remarks are provided in Section 4.

2. SUBJECTIVE EXPERIMENTAL DATABASE

A subjective experiment was designed to evaluate 4 encoding algorithms at bitrates ranging from 30 kbps to 80 kbps. Sixteen sign language stories told by a fluent signer at her natural signing pace were filmed in two different locations, an indoor studio with a static background and an outdoor location on a busy street having a significant amount of background activity.

A total of 16 fluent ASL users participated in the experiment, consisting of 7 male and 9 female participants, having an average of 28.4 years of experience with ASL. Participants provided their preferred language for communication: 8 reported English, 5 reported ASL, and 3 reported both English and ASL. The subjective experiment followed a single stimulus testing procedure. Following each test video, participants were asked three questions designed to evaluate their comprehension of the story, the intelligibility of the test video, and usability of the test video. Because of the nature of this intelligibility assessment task, no single story is viewed by the same participant twice, eliminating any possible learning effects.

The test set of videos evaluated in this study were generated using 4 different encoding algorithms each operating under both a high bitrate and low bitrate setting. Two different sets of rates were selected such that at each location, the most intelligible videos would be very easy to understand and the least intelligible video would be very difficult to understand. Outdoor videos were encoded at rates of 50 kbps and 80 kbps; indoor videos were encoded at rates of 30 kbps and 45 kbps. Each of the 4 encoding algorithms evaluated in this study operate within the H.264/AVC standard and provide rate control that meets an average target bitrate. Three sign language specific encoders and a general purpose video encoder were evaluated.

x264 Encoder An open-source H.264/AVC encoder, x264 [1], was selected because it offers rate-distortion performance similar to the JM reference encoder at speeds 50 times faster. The rate-distortion optimization in x264 effectively treats all macroblocks equally, minimizing the average MSE over the entire frame, subject to the rate constraint. Rate control is performed by adjusting the quantization parameter on a frame-by-frame basis [2].

Foveated Encoder In foveated video coding, the video frame is encoded with non-uniform, decreasing quality away from the the observer's point of fixation, attempting to match the visual acuity of the human visual system [3]. Because an ASL observer primarily gazes at the signer's face, the fixation point is automatically identified using skin segmentation and facial feature detection, and foveated processing is applied to generate a map of priority regions [4]. Relative to the location of the face, a foveation model assigns macroblocks to the priority regions, each region having a different quantization parameter, allowing blocks nearest to the face to be coded with more

bits than blocks farther away. Rate control is performed by adjusting the quantization parameter assigned to the face macroblocks (on which all the other quantization parameters depend) on a frame-by-frame basis.

ROI Encoders Two region-of-interest (ROI) encoding techniques allocate bits primarily to the face and hands of the signer [5]. Skin segmentation and face detection algorithms identify the macroblocks containing the signer's face and hands and provide a segmentation map to the ROI encoder. Given the segmentation map, different quantization parameters are selected for the macroblocks belonging to the face, hands, or background. Rate control is performed by adjusting the Lagrange multiplier on a frame-by-frame basis, which controls the quantization parameter selection for each region. These two ROI encoders are differentiated by the use of a temporal smoothing of the macroblock segmentation labels. Specifically, for the **spatial-temporal ROI encoder**, the face and hand labels are held constant for a duration of 1 second, e.g., if a macroblock is labeled as hand in frame 10, that macroblock will retain the hand label until frame 25, at 15 frames per second. The **spatial ROI encoder** simply uses the segmentation map provided for the current frame.

2.1. Verifying the necessity of z-scores

Analysis of variance (ANOVA) is used to identify statistically significant effects on subjective intelligibility and a significant effect was found for the preferred language ($F(2, 206) = 28.87, p < 0.01$). Participants who reported ASL as their preferred language responded with statistically significantly higher intelligibility ratings than both the group preferring English and the group preferring either ASL or English. Higher fluency is unlikely, as the group reporting both ASL and English has roughly the same mean years of experience with ASL. It is more likely that the ASL group was biased toward higher intelligibility ratings. This is a consequence of the increased desire an ASL user likely has for a mobile phone that offers video communication; making cell phone calls in their preferred language is currently unavailable for only this group of participants.

This phenomenon is confirmed by applying ANOVA to the subjective usability ratings. Similar to the results for intelligibility, the participant's preferred language is a significant effect and those preferring ASL responded significantly higher to the usability question, emphasizing their increased desire for such a technology. In order to eliminate this bias, the subjective intelligibility scores are converted to z-scores, which has the desired consequence of removing information about between subject differences [6]. In this experimental database, the use of z-scores is justified by a thorough analysis of the raw subjective ratings and z-scores necessarily remove a clear bias in a portion of participants. The z-scored intelligibility ratings are used for all further analysis.

2.2. Ranking encoders according to intelligibility

The ANOVA identifies a significant effect for the encoding algorithm ($F(3, 206) = 3.90, p < 0.01$) and encoding bitrate ($F(1, 206) = 31.46, p < 0.01$). For the high bitrate videos, there is no statistical difference in intelligibility between the x264, foveated, and spatial-temporal ROI encoding algorithms. The mean intelligibility z-scores for these encoders, presented in Table 1, correspond on average to raw intelligibility scores between "easy to understand" and "very easy to understand". Both the x264 encoder and the foveated encoder yield statistically significantly higher intelligibility than the spatial ROI encoder. The reduced performance of the spatial ROI encoder is a consequence of compression distortion artifacts in the signer's face and hands due to ROI segmentation errors. The spatial-temporal ROI encoder is less susceptible to segmentation errors because the region labels persist across time, eliminating any short duration segmentation errors. For encoding algorithms that rely heavily on region-based rate allocation, accurate segmentation is an important factor in the final subjective intelligibility.

Table 1. Mean and standard deviation for z-scored intelligibility ratings. Each average is computed over 32 individual ratings. A rating of "very difficult to understand" maps to an average z-score of -1.18 while a rating of "very easy to understand" maps to an average z-score of 1.13. In the high bitrate case, the x264, foveated, and temporally smoothed ROI encoders have statistically equivalent mean intelligibility scores. In the low bitrate case, the ASL-specific encoders result in statistically significantly higher intelligibility than the x264 encoder.

Bitrate	Algorithm	Mean	SD
High	x264	0.66	0.84
	Foveated	0.71	0.97
	Spatial-Temporal ROI	0.30	0.82
	Spatial ROI	-0.07	0.92
Low	x264	-0.90	0.70
	Foveated	0.01	0.80
	Spatial-Temporal ROI	-0.21	0.81
	Spatial ROI	-0.51	0.72

At low bitrates, all three of the ASL-specific encoding algorithms provide statistically significant improvements over x264 in mean intelligibility scores. The differences in intelligibility between the three ASL-specific encoders are not statistically significant, each having mean intelligibility z-scores corresponding to raw intelligibility ratings between "neither easy nor difficult" and "easy to understand". The mean intelligibility z-scores are provided in Table 1. The subjective experiment demonstrates that encoding algorithms designed specifically for ASL can provide statistically significant improvements in intelligibility over traditional, MSE-based encoding algorithms and can generate intelligible ASL video at bitrates as low as 30 kbps. More importantly, the experimental database affords a detailed analysis of distortions that adequately challenge objective estimators.

3. INFLUENCE OF SPACE AND TIME VARYING DISTORTIONS ON OBJECTIVE ESTIMATOR PERFORMANCE

While subjective testing is the most accurate method for evaluating distorted video, it is prohibitively costly. Objective estimators are desirable and are evaluated in terms of their ability to predict accurately subjective experimental data. The range of video content and distortion artifacts that occur in an experimental data set can bias conclusions regarding the performance of an objective estimator. In the context of objective intelligibility estimators, it is critical for the subjective test database to challenge the estimators by selecting source videos that vary in brightness and contrast and by applying distortions that vary non-uniformly both spatially, e.g., between the signer and the background, and temporally. This section analyzes the performance of 4 objective estimators on different subsets of the experimental data and demonstrates how the videos and distortions present can bias the conclusions made regarding an objective estimator.

Table 2. Objective estimator performance comparison. PSNR performs well for x264 coded videos at a single location, but fails to predict intelligibility when the locations are combined. RMS distortion contrast can predict intelligibility across the combined locations, but only for the compression artifacts. The objective intelligibility measure (OIM) offers improvements over both PSNR and RMS distortion contrast for videos containing spatially varying distortions. For the outdoor videos containing temporally varying distortions, the spatiotemporal OIM outperforms the spatial only OIM.

Source Videos	Measure	Data Set	RMSE	R
x264 Coded Spatially Uniform Distortions	PSNR	Indoor	0.96	0.899
		Outdoor	0.46	0.980
		Combined	2.18	0.094
	RMS Distortion Contrast	Indoor	1.01	0.886
		Outdoor	0.98	0.908
		Combined	1.46	0.743
	OIM Spatial Only	Indoor	1.01	0.886
		Outdoor	0.68	0.956
		Combined	1.14	0.853
	OIM	Indoor	0.98	0.894
		Outdoor	0.63	0.964
		Combined	1.06	0.875
ASL Coded Spatially Varying Distortions	PSNR	Indoor	1.54	0.090
		Outdoor	1.15	0.638
		Combined	1.49	0.056
	RMS Distortion Contrast	Indoor	1.53	0.141
		Outdoor	1.09	0.688
		Combined	1.47	0.190
	OIM Spatial Only	Indoor	1.34	0.499
		Outdoor	1.24	0.563
		Combined	1.29	0.509
	OIM	Indoor	1.36	0.480
		Outdoor	0.93	0.785
		Combined	1.23	0.569

Four objective estimators are considered for predicting intelligibility: PSNR, RMS distortion contrast, and an objective intelligibility measure (OIM), with and without appropriate temporal pooling [7]. For a video sequence, both PSNR and RMS distortion contrast are computed in each individual video frame and averaged across all frames. The OIM applies a region-based spatial pooling mechanism that allows for varying importance to be placed on the ROIs contained in sign language video and a temporal pooling mechanism that accounts for the temporal variations in distortions. The spatial only OIM includes only the region-based pooling mechanism.

The performance of an objective estimator is computed in terms of its prediction accuracy and linearity. For a given estimator, a linear fit is applied to map the objective estimate to the subjective ratings. Prediction linearity is measured using Pearson's linear correlation, R, and prediction accuracy is measured using the root mean squared error (RMSE) of the prediction residuals, after applying the linear fit. For this analysis, the experimental data is divided into two sets: the videos encoded using x264 and the videos encoded using the ASL-specific encoders. Within each of these 2 subsets, the prediction accuracy and linearity is computed for 3 additional subsets: the indoor and outdoor videos separately and the indoor and outdoor videos combined, resulting in 6 total data subsets. For all of the objective estimators evaluated, there is at least one experimental data set for which the estimator can accurately predict intelligibility.

PSNR performs well on uniform video content. When considering only the x264 coded set and separating the indoor and outdoor videos, PSNR provides an accurate estimate of subjective intelligibility, having low RMSE (0.96 and 0.46 for indoor and outdoor videos) and high linear correlation (0.899 and 0.980 for indoor and outdoor videos) similar to both RMS distortion contrast and the OIM, as illustrated by Table 2. One way in which energy-based error measures, such as PSNR, fail as objective estimators is with test videos containing varying brightness, because brighter videos yield higher MSE values at perceptually similar levels of distortion. Combining the indoor and outdoor videos into a single data set reveals the inability of PSNR to compare distortions across videos of varying contrast. The RMSE increases to 2.18 and the linear correlation decreases to $R = 0.094$, demonstrating that PSNR is virtually uncorrelated with the subjective intelligibility ratings when combining both indoor and outdoor video sets. Because it computes errors in contrast and not raw pixel values, RMS distortion contrast performs significantly better than PSNR on the combined data set, having an RMSE of 1.46 and correlation of $R = 0.743$.

RMS distortion contrast performs well for spatially and temporally uniform distortions. RMS distortion contrast provides an accurate estimate of subjective intelligibility for videos without spatially varying distortions, but fails to estimate intelligibility for ASL coded videos, as illustrated by Table 2. The OIM is the most accurate of the 3 objective estimators on the ASL coded set, having an RMSE of 1.23 and linear correlation of $R = 0.569$. The low correlation is primarily due to the difficulty in estimating intelligibility for the indoor videos. When comparing the ASL coded indoor and outdoor videos separately, the OIM performs well on the outdoor set, having RMSE equal to 0.93 and linear correlation of $R = 0.785$.

When the test data set includes only distortion artifacts resulting from compression, RMS distortion contrast and the OIM perform similarly and one could incorrectly conclude that applying knowledge of the ROI provides no benefit to an objective estimator. This behavior can be explained by the relationship between the global distortion and the region distortions. For simplicity in computation and discussion, MSE is used as the distortion measure in this comparison. For two sample videos (one indoor, one outdoor), the global MSE and the MSE in the face and hand regions are computed separately in each video frame, providing 3 vectors of per-frame MSEs having length equal to the number of frames. The correlation between the global MSE and each of the region MSEs is computed and presented in Table 3.

For videos coded using x264, which contain only compression artifacts, the global MSE is highly correlated with each of the region MSEs, having $R_{Face} = 0.875$, $R_{Hand} = 0.837$ for indoor videos and $R_{Face} = 0.775$, $R_{Hand} = 0.711$ for outdoor videos. Because of this, measuring distortion globally serves as an accurate predictor of the local distortions, which explains the high performance of RMS distortion contrast on just x264 coded videos. As a result of the high correlation, simply computing the global mean of the distortions can serve as a surrogate for perceptually valid region-based pooling. Knowing that an observer is extracting information from specific regions (e.g. face and hands) provides little improvement in prediction if the video is encoded without any region of interest approach. For videos encoded using the ASL-specific techniques, the distortions are very localized and global distortions are not correlated with the region distortions, having $R_{Face} = 0.028$, $R_{Hand} = -0.157$ for indoor videos and $R_{Face} = -0.154$, $R_{Hand} = 0.004$ for outdoor videos, coded using the spatial-temporal ROI encoder. In these ROI coded video cases, RMS distortion contrast fails to accurately estimate intelligibility. Objective estimators that incorporate region-based pooling, such as the OIM, are required to differentiate between

Table 3. Comparison of spatial and temporal variations in distortions. Correlation between global MSE and MSE in face and hand regions is given by R_{Face} and R_{Hand}, respectively. When considering only compression artifacts, such as those resulting from x264, the global MSE is highly correlated with distortion in the ROI. When the distortions vary spatially, global MSE and the ROI distortions are uncorrelated and a global distortion measure is unable to predict accurately the distortion in the ROI. TV measures the temporal variation of the distortions. Temporal distortions are relatively uniform for the indoor videos but vary substantially for the outdoor videos coded using the ROI encoders.

Location	Encoding Algorithm	R_{Face}	R_{Hand}	TV
Indoor	x264	0.875	0.837	3.5
	Foveated	0.261	0.714	5.0
	Spatial-Temporal ROI	0.028	-0.157	9.7
	Spatial ROI	0.069	-0.066	12.3
Outdoor	x264	0.775	0.711	9.0
	Foveated	0.166	-0.001	8.4
	Spatial-Temporal ROI	-0.154	0.004	233.1
	Spatial ROI	-0.096	0.042	268.2

distortions in the background and in the signer.

OIM performs well for spatially and temporally varying distortions. When the average distortion in a single frame is highly correlated with local distortions, the spatial pooling mechanism has little impact on the performance of a particular objective estimator. Similarly for temporal pooling, advanced pooling mechanisms are unlikely to improve the objective estimate when the distortions do not vary across frames. The temporal variation of the distortions in a sequence can be quantified by a measure based on between frame difference in distortions [8]. Videos in which the amount of distortion varies significantly between frames will yield a high measure of temporal variation. The outdoor video sequences coded using either of the ROI encoders have the largest measured temporal variation, having average temporal variation of $TV = 250.65$, while the indoor video sequences and the x264 coded sequences have very low temporal variation, having average temporal variation of 7.98, as summarized in Table 3.

When the temporal variation is low, as is the case with the x264 coded videos, advanced temporal pooling provides negligible improvement over a simple global average in terms of the prediction accuracy. In this case, the spatial only OIM performs similarly to the full spatiotemporal OIM, as illustrated in Table 2. However, when comparing the ASL coded outdoor data subset, which has the highest temporal variation, the spatial OIM, having RMSE of 1.24 and $R = 0.563$, performs significantly worse than the spatiotemporal OIM, having RMSE of 0.93 and $R = 0.785$.

4. CONCLUSION

Depending on the presence of varying video content and the homogeneity of the spatial and temporal distortions, several possible conclusions regarding the performance of objective intelligibility estimators can be incorrectly reached. Four objective estimators were evaluated in terms of their subjective intelligibility prediction accuracy. For each of the tested estimators, there exists a subset of videos on which the estimator can accurately predict subjective intelligibility. If the distortions present in a test set do not contain spatially or temporally localized errors, objective estimators that incorporate an ROI pooling mechanism cannot reliably improve the estimation accuracy, even in cases such as ASL video, when an ROI is unequivocally known to exist. This highlights the importance of both designing a diverse experimental data set and verifying that a proposed objective estimator is adequately challenged by the experimental data set on which it is tested.

In future work, the methodology of analyzing a subjective experimental database will be extended to other image and video quality databases. In the domain of quality assessment, many techniques have been proposed for improving objective quality estimators by incorporating visual attention data [9, 10]. Typically this is done by applying a spatially-varying weight to the errors prior to spatial pooling, wherein the weights are proportional to the relative importance of each region or pixel. Care must be taken to ensure that the distortions present in the subjective database effectively challenge the objective estimators. As demonstrated by this work, compression artifacts alone are insufficient for evaluating region-based pooling mechanisms.

5. REFERENCES

[1] x264. http://developers.videolan.org/x264.html.

[2] L. Merritt and R. Vanam, "Improved rate control and motion estimation for H.264 encoder," *Proc. IEEE International Conference on Image Processing*, vol. 5, 2007.

[3] P. Kortum, W. S. Geisler, B. E. Rogowitz, and J. P. Allebach, "Implementation of a foveated image coding system for image bandwidth reduction," *Human Vision and Electronic Imaging*, vol. 2657, pp. 350–360, Apr. 1996.

[4] 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*, no. 21, 2006, pp. 531–549.

[5] F. Ciaramello and S. Hemami, "Complexity constrained rate-distortion optimization of sign language video using an objective intelligibility metric," *Proc. SPIE Visual Communication and Image Processing*, vol. 6822, Jan. 2008.

[6] A. M. van Dijk and J. Martens, "Subjective quality assessment of compressed images," *Signal Processing*, vol. 58, no. 3, pp. 235–252, May 1997.

[7] F. M. Ciaramello and S. S. Hemami, "Quantifying the effect of disruptions to temporal coherence on the intelligibility of compressed American Sign Language video," in *Human Vision and Electronic Imaging XIV*, vol. 7240. San Jose, CA, USA: SPIE, Feb. 2009, pp. 72 400D–10.

[8] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, no. 2, pp. 253–265, 2009.

[9] ——, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in *Proc. IEEE International Conference on Image Processing*, vol. 2, 2007, pp. II – 169–II – 172.

[10] U. Engelke and H. Zepernick, "Framework for optimal region of interest–based quality assessment in wireless imaging," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011 005–13, 2010.