COMPRESSION FOR RECOGNITION AND CONTENT-BASED RETRIEVAL *

Antonio Ortega, Baltasar Beferull-Lozano, Naveen Srinivasamurthy and Hua Xie
Dept. of Electrical Engineering-Systems
Integrated Media Systems Center
University of Southern California Los Angeles, CA 90089-2564, USA
Tel: +1 213 740 2320; fax: +1 213 740 4651
e-mail: ortega,beferull,snaveen,huaxie@sipi.usc.edu

ABSTRACT

Most compression algorithms developed to date aim at achieving the best perceptual quality of the decoded media for the given rate. In this paper we consider several scenarios where the end user of the compressed data is not a human viewer or listener, but rather a known classifier or recognizer. Drawing from applications in speech recognition and image classification, as well as from simple examples, we discuss the new requirements that are imposed on the encoders under these circumstances. Our goal is to motivate the importance, and describe the associated design challenges, of achieving compression optimized for classification/recognition, rather than perceptual quality.

1 INTRODUCTION AND OUTLINE

Compression of multimedia data is an active research area and has led to the development of a series of standards such as JPEG or MPEG. In this paper, we will argue that novel compression techniques may be needed if the goal is to minimize the loss in recognition/classification performance due to compression, rather than minimizing the impact of coding on perceptual quality, which is the typical objective of standard techniques. We consider two example scenarios, namely, (i) wireless access to multimedia applications and (ii) storage and content based retrieval of multimedia data, in order to motivate the research described in this paper. In Section 2 we show that remote access to multimedia applications (e.g., those involving speech recognition) leads to distributed recognition/classification scenarios, where data is to be compressed before being transmitted in order to be processed by a remote recognition engine. An example is provided in the context of distributed speech recognition (Section 2.1) and the difficulties of designing compression for such distributed applications are illustrated with a simple example (Section 2.2). Then, in Section 3 we consider compression needs for multimedia databases. Given that context based retrieval needs to be supported, we motivate the potential benefits of non-traditional transforms, such as steerable decompositions (see Section 3.1). We then provide a simple example to demonstrate that a separate encoding of classification features can be useful in this application (see Section 3.2).

*This research has been funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152 and by NSF grant CCR-9804959.

2 DISTRIBUTED CLASSIFICATION

Wireless access to multimedia applications is limited by both power and bandwidth, so that compression is required before data is transmitted and computation at the mobile terminal should be kept at a minimum. Consider a scenario where a user is interacting with a remote application that employs speech recognition. Using a full fledged recognition engine at the transmitter may be too complex and instead it may be preferable to transmit compressed speech and have the recognition tasks performed remotely, a technique known as distributed speech recognition. In a distributed classification application, since encoder and classifier are physically apart, the classifier has to operate on decoded data. Thus, ideally, the encoder aims at quantizing more finely whatever information is more important for classification. This task is made difficult because, as seen in the speech case, the recognition engine tends to operate on larger dimensional data than the quantizer (e.g., a sequence of speech frames vs. individual frequency components.) We illustrate this problem by discussing a simple example where scalar quantization is used along with a vector based classifier.

2.1 Distributed speech recognition

We consider a scheme for distributed automatic speech recognition, where a Hidden Markov Model (HMM) based speech recognition system with a Mel Frequency Cepstral Coefficients (MFCC)[3] front end is used in the evaluation, and our goal is to achieve the best recognition performance for the given rate. At the client or mobile device the MFCCs are extracted from the speech and encoded with an algorithm [1] involving scalar quantization, linear prediction [4] and coefficient

pruning. Other approaches have been proposed (e.g., [5, 4]) that compress MFCCs but we have shown that good recognition performance can be achieved with simple coding techniques (i.e., no vector quantization) at low rates, while providing scalability, which will enable graceful degradation in recognition performance under time varying channel conditions.

The MFCCs are extracted after applying overlapping Hamming windows to the speech signal. Therefore we use one-step prediction so that each MFCC is predicted from the same MFCC in the previous frame. We then apply a simple uniform scalar quantization (USQ) to the prediction errors. Rate scalability can be achieved either through increases of the quantization stepsize, or through explicit pruning of certain coefficients (i.e., by only sending a subset of the MFCCs). Huffman coding is used to represent the quantization indices. Counting the cost of the MFCC computation (and without any significant code optimization) our approach is 3 times faster than applying the recognition algorithm itself.

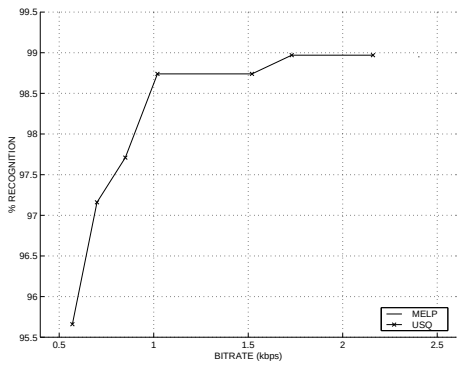

Figure 1: Recognition Performance for the USQ coder with different quantization stepsizes. The scalability of the encoders can be seen by the bitrate/recognition performance tradeoff.

Figure 1 summarizes our experiments performed on the TI46-Word digit database with HTK2.1. Details on the experimental setup can be found in [1]. The recognition performance for digit database when unquantized MFCCs were used was 99.79%. From Figure 1, it can be observed that the USQ technique, achieves recognition performance of 98.74% at a bitrate of 1.02 kbps We observe that the degradation with respect to the baseline in recognition performance is small, while there are substantial savings in the bitrate, with respect to using a standard coder. For example, using a MELP speech encoder at 2.4 kbps we achieved 98.85% recognition performance. Figure 1 also shows how scalable recognition performance can be achieved as the rate is reduced.

2.2 Coding for distributed classification

The previous section has given an example of how recognition performance of compressed data can be improved if the coding algorithm is designed to aim at recognition inaccuracy, rather than quantization distortion. In this section we provide a simple example of how one can operate an encoder while putting different emphasis on classification or distortion. It should be noted that previous work has considered the joint design of an encoder and a classifier where the system produces for each input a quantization index and a classification label [6, 7]. Here we consider a different problem, i.e., we encode the data first, and classification is applied separately. Moreover, we assume that we have no control over the classifier, while [6, 7] both design quantizer and classifier. Our task is further complicated because we consider cases (such as the one in the previous section) where a simple quantizer (e.g., scalar quantizers) needs to operate in conjunction with complex classifiers (e.g., based on vectors).

We consider an example where a two-dimensional classifier operates with data that has been scalar quantized separately in each dimension, i.e., the classifier takes as input vectors of quantized values. Our goal is then to design the scalar quantizers such that they the least effect on the classification performance for a given quantization distortion. However, since classification is vector-based and the quantizers are scalar we do not have exact knowledge of the effect of quantization design on classification. To address this we have developed a technique where scalar quantizer design in one dimension incorporates a measure of the average misclassification incurred over all the possible values taken by the other component of the vector.

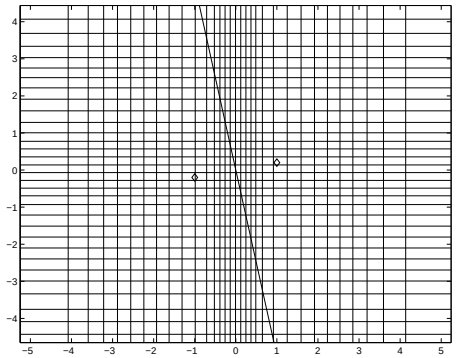

Figure 2: Quantization bins when the proposed algorithm is used to design the scalar quantizers. The diagonal line is the Bayes classifier boundary and the circles are the average values of the two Gaussian sources.

Results of our experiments are shown in Figs. 2 and 3 where a 2 dimensional classifier is used to separate two 2D Gaussian sources having same variance but different mean and where scalar quantization is used independently in each dimension. The best Bayes classifier for this data will assign an input vector to the class if the distance between the input and the mean of the class is minimal. Thus the classification boundary will be a line perpendicular to the line joining the mean of the two classes (see Fig. 2). We apply our design to both a scalar

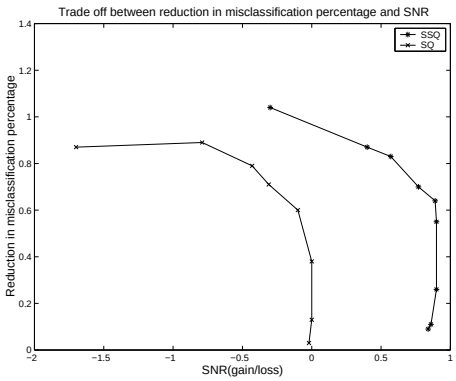

Figure 3: Trade-off between reduction in misclassification percentage and SNR for SQ and SSQ.

quantizer and to a sequential scalar quantizer (SSQ) [8], which provides lower distortion and lower classification error.

As can be seen in Fig. 2, our design selects smaller quantization bin sizes close to the optimal classification boundary. As in [6] the relative importance of distortion and misclassification is controlled by a Lagrange multiplier. This is illustrated in Fig. 3 where the various achievable points represent the distortion and classification performance for a fixed number of reproduction levels. As can be seen the best distortion performance corresponds to the worst classification performance.

3 CONTENT BASED RETRIEVAL

The second application we consider is content-based access to databases containing large amounts of multimedia data, where text-based indexing is not sufficient and we assume content features are available. Consider for example an image database, where, obviously, images are stored in a compressed format. Assume a user, remotely or locally, sends a query to the database. This query is based on the image content (e.g., color content, presence of a specific texture, etc.) and thus the content of each of the images in the database will have to be analyzed. One could decompress each of the images to extract the relevant features but that would be clearly impractical. A popular alternative is to extract the feature information from the compressed domain. This approach is certainly simpler but it has two major drawbacks. First, not all features of interest may be extracted from the compressed domain (e.g., due to the fact that images are broken into blocks in a DCT based coder.) A second drawback is that the amount of data to be manipulated is still large (even after compression) and therefore disk I/O may become a significant overhead. Here we consider two alternative approaches based on (i) using alternative transforms that can preserve features of interest and (ii) storing and compressing the feature vectors required by each class of queries.

3.1 Steerable transforms

If certain features of interest are not preserved by standard critically sampled transforms such as the discrete wavelet transform it is possible to consider overcomplete transforms instead. As an example we consider the use of steerable representations in applications where it is necessary to extract many different features and therefore selectivity in orientation, rotation-invariance and approximate shift-invariance are desirable. Since these properties are not available with critically sampled transforms our goal is to define compression algorithms for steerable decompositions. While steerable decompositions are overcomplete and thus increase the amount of data to be compressed, they also present some properties that can be exploited for more efficient compression.

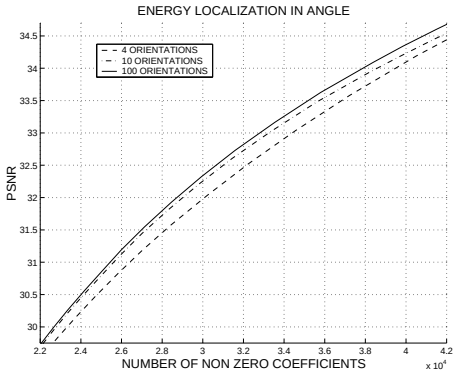

Figure 4: Energy localization in angle. Observe that with angular oversampling the same energy is concentrated in fewer coefficients

A filter (or function) $h(x, y)$ is called steerable if transformed versions of this filter can be expressed using linear combinations of a fixed set of basis filters. Following the analytical approach to steerability developed mainly by Freeman and Adelson [9] and later extended by Simoncelli [10], one can design filters, with certain restricted analytical forms, which are polar separable in the Fourier domain and where the steerable basis filters are steered versions of the steerable filter itself.

Thus, if a filter $H(\omega_x, \omega_y)$ is polar separable in the Fourier domain, then $H(\omega_x, \omega_y) = B(r)G(\phi)$, and steerability of $H(\omega_x, \omega_y)$ is equivalent to shiftability of $G(\phi)$. Given $G(\phi)$, one can find a set of N angles $\phi_0, \phi_1, \ldots, \phi_{N-1}$ and a set of N interpolation functions $\{b_0(\phi), b_1(\phi), \ldots, b_{N-1}(\phi)\}$ such that the following is satisfied:

$$G(\phi - \alpha) = \sum_{n=0}^{N-1} b_n(\alpha) G(\phi - \phi_n) \quad \forall \alpha \qquad (1)$$

This enables one to, starting with N basis steerable filters or angles, derive filtered values at any other angular position. This can be useful for classification but results in a significant oversampling factor, e.g., approximately $\frac{4N}{3}$ when using a recursive pyramidal structure

on the radial part [2]. We have shown that angular oversampling (i.e., using more angles than the minimum N required for reconstruction) results in increased energy localization. This is illustrated in Figure 4 where we can see that as the oversampling increases the number of coefficients needed to achieve a given PSNR decrease. Our ongoing work aims at exploiting this and other properties to provide an efficient coding for steerable transforms (refer to [2] for further details).

3.2 Compression of texture features

In content-based image/video retrieval systems, multiple visual features such as texture, color, shape and motion are extracted automatically and used as indexing keys. When databases become large it may be impractical to extract these features on the fly from the compressed images. In this section we show a simple example that demonstrates that storing the feature set in compressed format along with the image and video data can lead to significant reductions in processing time, without reduction in classification performance.

To illustrate this idea, in our experiments we compare the performance of the various techniques using the wavelet based texture classification technique of [11], with 10 512x512 texture images from the Brodatz's texture album. As test images, we used randomly selected smaller subimages of the original texture images. Four-level Dyadic wavelet decomposition is performed on the sample images using a 16-tap Daubechies filter. The feature vector for a given image is composed of the averaged absolute energy of each wavelet subband, i.e., each component of the vector corresponds to one subband [11]. We use the simplified Mahalanobis distance to compare feature vectors [11].

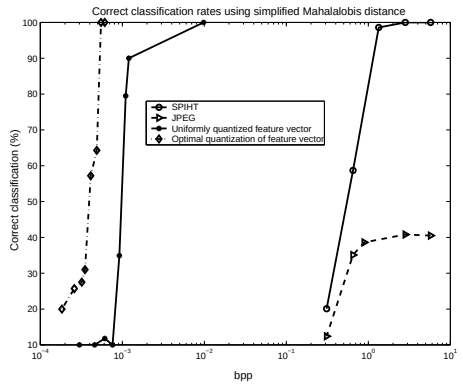

Figure 5: Comparison of classification performance using different quantization approaches.

We designed scalar quantizers to encode feature vectors, i.e., the set of average energies in each subband. We compare quantizers that have been optimized based on the training set with uniform quantizers where the same stepsize is used for each of the elements of the feature vector. In both cases 4 bits per subband are used. The efficiency of this approach can be observed Figure 5 where we see that the same classification rate can be achieved with close to three orders of magnitude fewer bits than with SPIHT [12]. This indicates that it may be advantageous to store these compressed features separately and to use them directly in the query process, instead of extracting the feature vectors from the compressed images.

References

[1] N. Srinivasamurthy, A. Ortega, Q. Zhu, and A. Alwan, "Towards efficient and scalable speech compression schemes for robust speech recognition applications," in *IEEE Intl. Conf. on Multimedia 2000*, July 2000.

[2] B. Beferull-Lozano and A. Ortega, "Coding techniques for oversampled steerable transforms," *Asilomar Conf. on signals, systems and computers*, 1999.

[3] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, 1980.

[4] G. N. Ramaswamy and P. S. Gopalakrishnan, "Compression of acoustic features for speech recognition in network environments," in *IEEE ICASSP 1998*, pp. 977–980, 1998.

[5] V. V. Digalakis and L. G. Neumeyer, "Quantization of cepstral parameters for speech recognition over the world wide web," *IEEE Journal on Selected Areas in Communication*, vol. 17, pp. 82–90, January 1999.

[6] K. O. Perlmutter, S. M. Perlmutter, R. M. Gray, R. A. Olsen, and K. L. Oehler, "Bayes risk weighted vector quantization with posterior estimation for image compression and classification.," *IEEE Trans. Image Proc.*, vol. 5, pp. 347–360, February 1996.

[7] J. Li, R. M. Gray, and R. A. Olshen, "Joint image compression and classification with vector quantization and a two dimensional hidden markov model," in *DCC 1999*, pp. 23–32, March 1999.

[8] R. Balasubramanian, C. Bouman, and J. Allebach, "Sequential scalar quantization of vectors: An analysis," *IEEE Tr. Image Proc.*, vol. 4, pp. 1282–1295, Sept. 1995.

[9] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 13, no. 9, pp. 891–906, 1991.

[10] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *IEEE Trans. on Info. Th.*, vol. 38, no. 2, pp. 587–602, 1992.

[11] T. Chang and C.-C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Trans. on Image Proc.*, vol. 2, pp. 429–441, Oct. 1993.

[12] A. Said and W. A. Pearlman, "A new fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. on Circ. and Syst. for Video Tech.*, vol. 6, pp. 243–250, June. 1996.