A MODEL-BASED ENHANCED APPROACH TO DISTRIBUTED VIDEO CODING[1]

Xavier Artigas, Luis Torres

Technical University of Catalonia, Barcelona, Spain
{xavi, luis}@gps.tsc.upc.edu

ABSTRACT

Distributed Source Coding (DSC), has been recently attracting a lot of attention from the video coding community, since it could prove very well suited for some applications where low-complexity encoders are a must. In order to understand how DSC can help these new applications, this paper quickly summarizes its theoretical bases and reviews the current state of the art for the particular but important case of Distributed Video Coding (DVC). Furthermore, a model-based enhanced approach to the DVC problem is presented that can be used when focusing on videoconference applications. The main novelty of this approach is the improvement of the side-information by means of a 3D face model, and the introduction of iterative decoding which improves the quality of the decoded sequence.

1. INTRODUCTION

Video coding research and standardization have been adopting until now a video coding paradigm where it is the task of the encoder to explore the source statistics, leading to a complexity balance where complex encoders interact with simpler decoders. This paradigm is strongly dominated and determined by applications such as broadcasting, video on demand, and video streaming. Distributed Video Coding (a particularization of Distributed Source Coding) adopts a completely different coding paradigm by giving the decoder the task to exploit the source statistics to achieve efficient compression. This change of paradigm also moves the encoder-decoder complexity balance, theoretically allowing the provision of efficient compression solutions with simple encoders and complex decoders. This coding paradigm is particularly adequate to emerging applications such as wireless video cameras and wireless low-power surveillance networks, disposable video cameras, medical applications, sensor networks, multi-view image acquisition, networked camcorders, etc., where low complexity encoders are a must because memory, computation, and energy are scarce.

However, even though the theoretical bases for Distributed Source Coding were set thirty years ago with the work by Slepian & Wolf [1] (for the lossless case) and Wyner & Ziv [2] (for the lossy case), it has been only recently that research on the topic has taken a new momentum. This research has been encouraged by the rise of some new applications, and has been leaded mainly by Ramchandran et al. at Berkeley [3] and Girod et al. at Stanford [4]. An excellent review of other works can be found in [4].

Although Distributed Source Coding can be used in other areas, like Robust Channel Transmission, this paper focuses purely on the aspects related to compression using low-complexity encoders.

The main innovation presented here consists in the usage of a 3D model of the sequence being transmitted to try to improve the quality of the motion-compensated estimation performed at the decoder. This restricts the area of application of this codec, but also, by focusing in a particular application (like videoconferencing), the decoder has extra knowledge that can be used to help the decoding process.

Section 2 introduces Slepian and Wolf's and Wyner and Ziv's theorems for Distributed Source Coding. Section 3 then summarizes the approach followed by Stanford, for the particular case of Distributed Video Coding. Next, Section 4 presents the novel model based approach. Simulation results and comparison to current state-of-the-art are given in Section 5, and, finally, Section 6 extracts some conclusions.

2. THEORETICAL FOUNDATIONS

2.1. The Slepian-Wolf theorem (lossless source coding)

It is a well known fact that the minimum lossless rate at which a signal X can be transmitted is $H(X)$, the signal's entropy. It is also well known that if two statistically dependent signals X and Y are to be transmitted, the best thing that can be done is to encode them together, in order to exploit the statistical dependencies, and that the minimum lossless rate is then $H(X, Y)$, their joint entropy. Slepian and Wolf showed in 1976 [1] that this lower

[1] ACKNOWLEDGMENT: The work presented was developed within VISNET, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 programme.

bound for the lossless joint transmission rate is also achievable when the signals X and Y are encoded *separately*, provided that some conditions are fulfilled. That is, when the encoder for X does not have access to Y, and vice versa. A codec that exploits this theorem is called a *Slepian-Wolf codec*.

In a nutshell, what the Slepian-Wolf theorem states is that, two statistically dependent signals X and Y, can be *separately encoded* and still be jointly recovered at the receiver with an arbitrarily small error probability as long as the following conditions are met [1]:

$$R_X \geq H(X|Y) \quad (1)$$

$$R_Y \geq H(Y|X) \quad (2)$$

$$R_X + R_Y \geq H(X,Y) \quad (3)$$

Where R_X and R_Y are the transmission rates of X and Y respectively, $H(X,Y)$ is their joint entropy, and $H(X|Y)$ and $H(Y|X)$ are their conditional entropies. The error probability can be made smaller by enlarging the frame length (this is, sending data bits in packets and jointly decoding every bit in a packet), approaching zero as the frame length approaches infinity.

It is the ability of encoding X and Y separately that makes Distributed Source Coding so attractive, because encoders for separate signals do not have to search for inter-correlations among signals, and therefore require fewer computations. Note that, in order to correctly decode the transmitted signals, these inter-correlations still have to be found, but this is now done in the decoder. On an implementation context, this means that the complexity of the coder is transferred to the decoder.

2.2. The Wyner-Ziv theorem (lossy source coding)

Three years later, the work from Wyner and Ziv [2] extended the work by Slepian and Wolf [1] by studying the lossy case in the same scenario, where signals X and Y are statistically dependent. Y is transmitted at a rate equal to its entropy (Y is then called *Side Information*) and what needs to be found is the minimum transmission rate for X that introduces no more than a certain distortion D (for some distortion measure). The Wyner-Ziv theorem introduces the Wyner-Ziv rate-distortion function, which is the lowest bound for R_X, the aforementioned transmission rate. A codec that intends to separately encode signals X and Y while jointly decoding them, but does not aim at recovering them perfectly (i.e. it expects some distortion D in the reconstruction) is called a *Wyner-Ziv codec*, and it can use the Wyner-Ziv rate-distortion function as a bound for its efficiency.

3. CURRENT APPROACHES

The proofs of the above theorems are asymptotical and non-constructive, meaning that the implementation of a codec based on them is not straightforward, and, as a consequence, different approaches are currently being explored. The two approaches that have had more continued effort in the field of video coding are the one by Ramchandran et al., at Berkeley [3], and the one by Girod et al., at Stanford [4][5][6], which is reviewed next.

Girod et al. have developed a distributed video codec based on techniques borrowed from channel coding [4]. The idea is to treat Y (the side information) as a *noisy version* of X (the main signal). Then Y may be sent using a *conventional coding approach* while X is sent at a rate lower than its entropy (hence introducing loss). The original signal X does not actually need to be sent (since the receiver already has a *noisy* version of it), instead, only the necessary data to recover it from Y is transmitted. The way in which this data is generated is taken from channel coding theory, under the form of parity bits. Therefore, the overall process in this approach is as follows: Y is conventionally encoded (by means of entropy coding or intra-frame coding, for example) and transmitted; parity bits are calculated for X and transmitted; the receiver then applies those parity bits to Y to recover X.

In [4], side information Y is generated as follows: key pictures are transmitted using intra-frame coding and the receiver uses them to generate an estimate of the missing pictures X, using motion-compensated temporal interpolation (MCTI). Parity data is generated using Turbo Codes, and, in order to attain higher compression, the amount of parity bits that are sent varies, depending on the changing statistics between X and Y. This way, only the strictly minimum necessary information is sent. This is achieved by using a return channel: the encoder initially supplies a small number of parity bits, and the decoder is allowed to ask for more parity bits when decoding does not succeed with sufficient reliability.

While still not reaching the performance of state-of-the-art inter-frame coding, the system is reported to perform 10-12 dB better than H.263+ intra-frame coding. It is also shown in [4] that the better the estimation of the missing pictures (MCTI), the better the performance (better estimation implies higher correlation between X and Y, which means that fewer parity bits are required to successfully decode the original picture).

4. MODEL BASED DISTRIBUTED VIDEO CODING

The main concept of this paper is to combine model-based coding with a distributed approach. Traditional model-based coding generates a model of the picture being encoded (or adapts a pre-existing model), and transmits the model parameters to the decoder. In a distributed approach, the goal is to perform this model fitting at the decoder, so the encoder is kept as simple as possible, providing the benefits reviewed in the introduction. This scheme can be seen as the distributed version of traditional model-based coding, or as a model-based enhancement for distributed coding. Either way, it is expected to improve the individual usage of these two techniques. This paper focuses on videoconferencing applications, so most of the time the pictures being encoded will contain a human head in the foreground; therefore, only a generic deformable model of a human head is needed.

The objective that is being pursued with this approach is twofold; firstly, the model improves the side information, and secondly, it allows iterative decoding, which can successively improve the quality of the decoded images.

As seen in the previous sections, the better estimation of the picture being decoded the receiver has, the less information the encoder will need to transmit. The first objective of this paper is to present a mechanism to improve this estimation, by taking advantage of the *a priori* knowledge the decoder may have about the picture. It is intuitive that if the transmitted image is a human head, and the decoder knows it, then only information regarding this particular head needs to be transmitted. A lot of information which is common to all human heads can be safely omitted from the transmission, since the decoder already has it, in the form of a generic model. This principle, which has already been used in previous model-based approaches to conventional (non-distributed) coding, will be used here in a slightly different way.

Using Girod's approach [4] as a starting point, the process is summarized as follows: Some frames are intra encoded and decoded. A 3D model of a human head is adjusted so the main features of the model (eyes, nose, mouth) roughly correspond to the received intra frames. The rest of the frames, called Wyner-Ziv frames, use the DVC approach, and therefore require the decoder to generate an estimation to be used as side information. This estimation has been produced in this way: MCTI is used in a first step, and will be used to synthesize the background. Neighboring intra frames are then used as references; the parameters used to adjust the 3D model to the reference frames are interpolated, to generate the model for the frame being decoded. Finally, the interpolated model is textured using the reference frames (Figure 1): every group of pixels under a model triangle in the reference frame is warped (shrunk, stretched, …) to fit the corresponding triangle in the target frame (the frame being decoded). When more than one reference frame is used, all contributions to the final frame are averaged. The textured model is then drawn on top of the MCTI estimation to generate the final estimation to be used as side information. The whole process is depicted in Figure 2 (The distributed encoder and decoder follow Girod's approach [4], the *parameter extraction* module finds the model parameters that suit the input image the best, and the *parameter application* module is the renderer, which generates a synthetic image based on the 3D model and the estimated parameters).

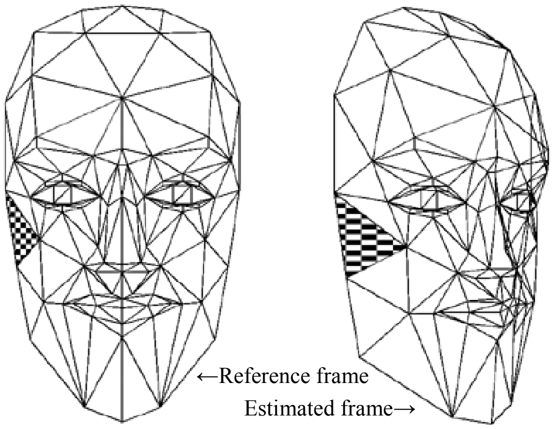

Figure 1 Pixel triangles in the reference frame are warped to build the estimated frame.

The second objective of this paper is to show that the aforementioned process to generate the side information can be iterated. It works as follows: Once the parity bits have been applied to the side information generated as explained above, a *partially decoded frame* is produced, which contains information about the original frame not present in the reference frames used for the interpolation. At this point, the model for the frame being decoded can be adjusted to this partially decoded frame, to correct for possible differences between the original frame and the interpolation. This adjusted model can then be used to generate a second side information, and the parity bits re-applied (Figure 2). This process can clearly be iterated many times, and each run should produce more accurate results as the estimation of the model parameters becomes more accurate.

As can be deduced from the explanations above, final image quality is improved not by augmenting the number of transmitted parity bits, but by improving the quality of the estimation. This extra quality comes both from the a priori knowledge the decoder has about the picture and the progressive extraction of the information contained in the parity bits.

5. SIMULATION RESULTS

Since at the current stage of this research the objective is to check the validity of the distributed model-based approach, the model is being adjusted manually. Every frame is fitted to the partially decoded frame using interactive software, and requires user intervention.

The proposed model-based system has been built on top of a DVC codec following the approach in [4]. A return channel is used, and the error probability required to decide if more parity bits are needed is assumed

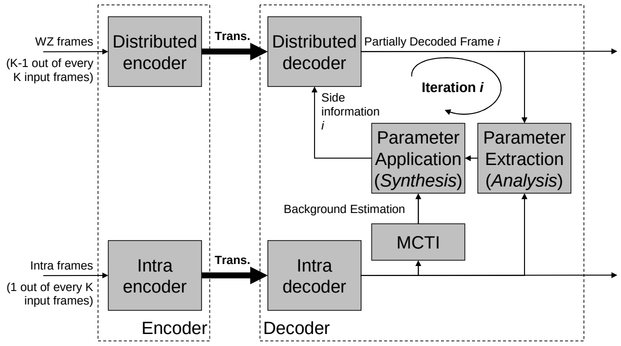

Figure 2 Scheme for the proposed model-based distributed video codec.

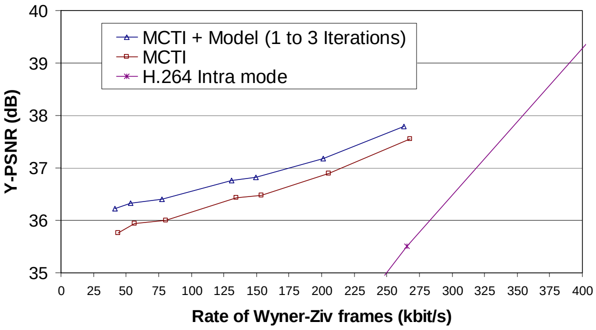

Figure 3 Rate-distortion comparison of the model-based approach with MCTI and H.264 in intra mode.

available at the decoder (ideal error detection). 100 frames of the Foreman sequence in QCIF format at 30 Hz. have been coded. Even frames were intra coded and decoded while odd frames, the Wyner-Ziv frames, used the proposed codec. The MCTI uses symmetrical bidirectional block matching and overlapped block motion compensation. The 3D model used is a slightly modified version of the Candide model [7], using the two neighboring frames as reference frames for the texturing process.

Rate-distortion plots are shown in Figure 3, along with H.264 in intra mode, for comparison purposes with a codec of similar encoder complexity. The rate axis is the rate of the Wyner-Ziv frames, that is, the frames that are not intra coded. It can be seen that the addition of the model-based side information generation plus the iterative refinement of the model parameters add from 0.25 to 0.45 dB to the PSNR of the sequence decoded using only MCTI as side information. Since the manual adjustment of the parameters is very good, only one to three iterations are required to reach a stable output.

Intermediate steps in the decoding of a particularly difficult frame (number 93) are shown in Figure 4. For this frame, the MCTI gives a very low PSNR (a), which is partially corrected by the model (b). It is interesting to note that the model does not suffer (and will never suffer) from the triple-eye effect, although it comes with its own artifacts, like the seams at the perimeter of the face. The parity bits sent by the encoder correct (to a certain amount) the remaining errors (c).

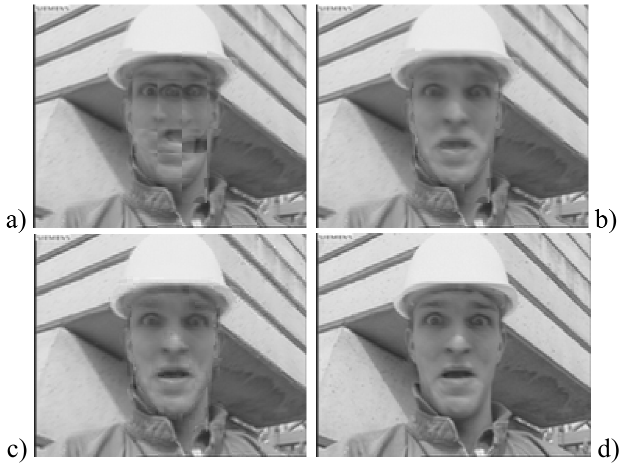

Figure 4 Frame 93: **a)** MCTI **b)** MCTI with the 3D model rendered on top **c)** decoded picture (MCTI + model + parity bits) **d)** original picture

6. CONCLUSION

It has been shown that, in a distributed coding approach, by using a 3D model, a priori information about the sequence being transmitted can be incorporated into the decoding process, increasing the quality of the produced pictures.

Moreover, refining the model parameters by adjusting them to the partially decoded frames allows iterative decoding and progressive enhancement of the decoded frames.

7. REFERENCES

[1] D. Slepian and J. Wolf, “Noiseless coding of correlated information sources”, *IEEE Trans. Inform. Theory*, vol. 19 pp. 471-480, July 1973.

[2] A. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder”, *IEEE Trans. Inform. Theory*, vol. 22, pp. 1-10, January 1976.

[3] R. Puri and K. Ramch andran. “PRISM: A new robust video coding architecture based on distributed compression principles”. *Proc. of 40th Allerton Conf. on Comm., Control, and Computing*, Allerton, IL, Oct. 2002.

[4] B. Girod, A. Aaron, S. Rane and D. Rebollo-Monedero, “Distributed video coding”, P*roc. of the IEEE*, vol. 93, no. 1, January 2005

[5] A. Aaron, E. Setton, B. Girod, “Towards practical wyner-ziv coding of video”, *Proc. ICIP 2003, Volume: 3 , 14-17 Sept. 2003*.

[6] A. Aaron, B. Girod, “Compression with side information using turbo codes”, *Proceedings DCC 2002 , 2-4 April 2002* Pages: 252 - 261

[7] http://www.bk.isy.liu.se/candide/