CONF-960672--1

UCRL-JC-122389
PREPRINT

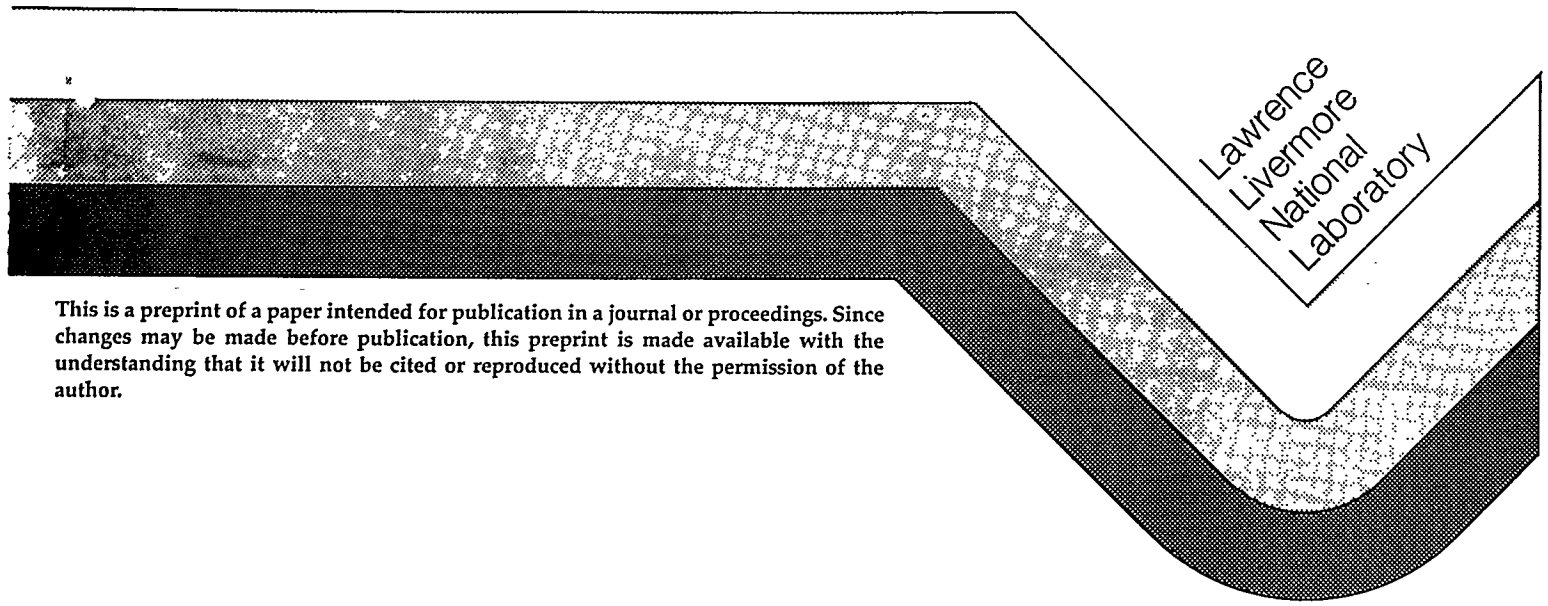
"Segmentation-based Video Coding"

Martin Lades
Institute for Scientific Computing Research
Lawrence Livermore National Laboratory
Livermore, CA

Yiu-fai Wong, Qi Li
University of Texas
Division of Engineering
San Antonio, Texas

This paper was prepared for submission to
IEEE Computer Society Conference on
Computer Vision Pattern Recognition '96
San Francisco, CA June 16 - 20, 1996

October, 1995



Segmentation-based Video Coding

Martin Lades
Institute for Scientific Computing Research
Lawrence Livermore National Laboratory
L-416, Livermore, CA 94551
hml@llnl.gov

Yiu-fai Wong and Qi Li
University of Texas
Division of Engineering
San Antonio, TX 78249
iwong@runner.jpl.utsa.edu

October 24, 1995

Abstract

Low bit rate video coding is gaining attention through a current wave of consumer oriented multimedia applications which aim, e.g., for video conferencing over telephone lines or for wireless communication. In this work we describe a new segmentation-based approach to video coding which belongs to a class of paradigms appearing very promising among the various proposed methods. Our method uses a nonlinear measure of local variance to identify the smooth areas in an image in a more indicative and robust fashion: First, the local minima in the variance image are identified. These minima then serve as seeds for the segmentation of the image with a watershed algorithm. Regions and their contours are extracted. Motion compensation is used to predict the change of regions between previous frames and the current frame. The error signal is then quantized. To reduce the number of regions and contours, we use the motion information to assist the segmentation process, to merge regions, resulting in a further reduction in bit rate. Our scheme has been tested and good results have been obtained.

Categories: Applications
Keywords: Video coding, segmentation, watershed.

Segmentation-Based Video Coding

Abstract

Low bit rate video coding is gaining attention through a current wave of consumer oriented multimedia applications which aim, e.g., for video conferencing over telephone lines or for wireless communication. In this work we describe a new segmentation-based approach to video coding which belongs to a class of paradigms appearing very promising among the various proposed methods. Our method uses a nonlinear measure of local variance to identify the smooth areas in an image in a more indicative and robust fashion: First, the local minima in the variance image are identified. These minima then serve as seeds for the segmentation of the image with a watershed algorithm. Regions and their contours are extracted. Motion compensation is used to predict the change of regions between previous frames and the current frame. The error signal is then quantized. To reduce the number of regions and contours, we use the motion information to assist the segmentation process, to merge regions, resulting in a further reduction in bit rate. Our scheme has been tested and good results have been obtained.

Summary

1. What is the original contribution of this work?

We have a new segmentation algorithm based on a robust measure of local activity in the image. Furthermore we use motion information to assist the segmentation of an image sequence. This results in the reduction of the number of regions an simplifies the resulting contours.

2. Why should this contribution be considered important?

Low bitrate video coding is critical for the emerging multimedia and wireless communication markets.

3. What is the most closely related work by others and how does this work differ?

Most closely related is the work by Salembier, Ebrahimi, Gu, and Kunt. We use motion information to merge regions and simplify the segmentation results.

4. How can other researchers make use of the results of this work?

We can promote the awareness of low bitrate to the computer vision community and use the work as basis for future model-based encoding approaches.

This work has not been previously presented at or submitted to other conferences, workshops, or journals.

1 Introduction

New application areas opened by the merging of computer, television and telephone are creating a huge demand for data compression due to their transmission bandwidth requirements. Applications in video conferencing or mobile and personal communications demand in addition to the adoption of coding standards such as MPEG-1 and MPEG-2 a new generation of coding techniques that have been coined "very low bitrate video coding." In very low bitrate video, we target a data bandwidth not exceeding 64 kb/s. Because the requirements and standards are still to be determined [1], many encoding approaches have been proposed [3]. The segmentation-based approach to coding is popular and shows good results [2, 3]. In our approach segmentation techniques are used to extract objects of arbitrary shape as video primitives. The shape and texture of the video primitives is then encoded and transmitted.

Needless to say that segmentation is an important building block in a segmentation-based video encoding approach. It is important that this segmentation is computed in a simple and efficient manner. Segmentation based on morphology is one popular approach and has been used in [2] where segmentation proceeds from coarse level to fine level. It is known that morphological operations are expensive when the structure elements are large. Thus, the method proposed in [2] is computationally expensive in those cases. In our work, we propose a hierarchical method for image segmentation. This method is shown to work fast and to obtain good segmentation results.

Another important feature in any video coding system is the use of temporal correlation for compression. In a segmentation-based approach, motion estimation is carried out for each of the segmented regions. That is, one computes the region-to-region correspondence, instead of matching blocks as in earlier encoding standards. However, an object is likely to be broken up into many smaller regions in segmentation. Thus, one needs to encode many regions and contours. This is expensive in terms of the number of bits required. We propose a merging scheme that merges regions based on both spatial and motion similarity. This differentiates our work from that in [2] where the authors merge regions based solely on spatial similarity. The advantage of our approach is that objects containing intra-object spatial dissimilarity will not be broken into many regions, resulting in a huge saving in the encoded data amount.

This contribution is organized as follows: We first describe our segmentation method in

Section 2. Then we describe our coding scheme in Section 3 and present some results in Section 4, followed by a discussion.

2 Segmentation

Watershed [5] is a powerful idea for segmentation. It is typically applied to gradient or gradient-related measures from an image. The difficulty is that the usual gradient measures are sensitive to noise. Thus, oversegmentation occurs and too many regions are generated. Elaborate nonlinear methods [6] are used to clean out the noise before one can actually use the regions.

It is important to pick the right seeds to initiate the watershed process. One should not have multiple seeds in the same region because that will break a region into smaller parts, leading to oversegmentation. To get good initial seeds, we utilize a new measure of local variance, which is very robust with respect to noise. First, let us look at the standard measure of local variance:

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \bar{x})^2 \tag{1}$$

where $\bar{x}_i$ is the local mean. Because the weights are uniform, this measure is not robust to noise. Thus, if one uses the local minima of variance to extract the seeds, there will be too many of them.

Suppose that one can adapt the weights so that noisy points have smaller weights, then we will have a variance that is more robust. This is exactly what the clustering filter can do [4]. The clustering filter computes an output that is a nonlinearly weighted sum of the neighboring pixels. The weight decreases with the distance between the pixel locations and distance between the filter output and the pixel value. Thus, noisy pixels that are different from the filter output will be weighed much less. Thus, we have a measure called "energy":

$$E(i) = \sum_j (y_i - x_j)^2 P_j. \tag{2}$$

where $P_j = e^{-\alpha(i-j)^2 - \beta(y_i - x_j)^2}/Z$ and Z is a normalizing constant. P_j suppresses the contribution of noisy pixels to the energy measure in Eq. (2). If we compute this measure for every pixel, we generate the so-called energy image.

Although we have a more robust measure, we will still suffer from the excessively large number of local minima in a large flat region. The solution we devised is to use multiresolution techniques to reduce the energy image to a lower resolution. Then on the reduced image, the number of local minima is likely to be small. We have found that this method leads to a reasonable number of seeds for watershed and good results are obtained. In the multiresolution method that we use here, we just use simple averaging, followed by subsampling.

Once the local minima of variance are selected the watershed algorithm [5] partitions the whole image into regions.

3 Coding Scheme

Let us discuss how to encode the first frame and frames within a sequence with our method. After an image is segmented, the next task is to encode the regions. A region has two components: its shape and the texture inside. The shape is encoded by contour coding while the texture is coded by polynomial approximation. The contour in each region needs to be cleaned up to avoid high cost associated with coding convoluted contours. We developed a nonlinear operation which can smooth contours.

For subsequent frames, temporal correlation is exploited through motion compensation. The idea is that if we can compute the movements of the segmented objects, we can predict the current frame from previous frames. The predicted image can be encoded by the previous frame and the motion information only, thus resulting in a substantially reduced number of bit needed for encoding.

During the segmentation process, an object may consist of several regions. This results in a larger number of regions and contours then we need to encode the frame. Thus, it is desirable to combine such regions encoding a single object into one region. If the regions belong to the same object, then their motion information tends to be similar. It is therefore necessary to merge regions based on their motion information. We have devised a scheme where merging occurs if the motion vectors for two neighboring regions are less than a threshold. This merging continues until no neighboring regions can be merged.

After motion compensation, high quality pictures also demand the encoding and trans-

mission of the difference image. In our method, a one dimensional signal is formed from the rows of each region. The signal is then encoded by a wavelet transform [7] and scalar quantization of the coefficients. The contour in each region is then chain-coded. Figure 1 shows the block diagram for our coding method.

Before the information about the data is sent to a channel coder, the motion information, the contour and the wavelets coefficients are further compressed by arithmetic coding.

4 Results

The scheme has been implemented and tested on the *Miss America* sequence. Generally, good results have been obtained. Figure 2 shows the compression ratio for 98 frames in the sequence. We obtained a compression ratio of around 350:1 for most frames. Figure 3 shows the corresponding SNR ratios. The picture quality is better than achieved by block-based coding methods. No blockiness effect is observed. Texture areas are also well-preserved.

5 Discussion

In this report, we have discussed a segmentation-based video coding scheme that is aimed towards very low bitrate applications. The segmentation scheme is based on multiresolution ideas and a robust measure of local variance. A watershed algorithm is used to segment the image into various regions. Temporal correlation is exploited to predict the motion of the segmented regions in successive frames. The difference signals between the original and the predicted frames are encoded by a wavelet expansion. Finally the necessary bits required to specify the regions, contours, motion and wavelet coefficients are further compressed by arithmetic coding.

The preliminary results have been very encouraging. The basic work can definitely be improved on in the future. We are currently working on improving the segmentation method and coding efficiency of the various modules.

References

[1] Y.Q. Zhang, "Very Low Bit Rate Video Coding Standards," *Proc. VCIP*, 1016–1023, 1995.

[2] P.Salembier, et al, "Region-based video coding using mathematical morphology," *Proc. IEEE*, 83, 843–857, 1995.

[3] T. Ebrahimi, E. Reusens and W. Li, "New trends in very low bitrate video coding," *Proc. IEEE*, 83, 877–891, 1995.

[4] Yiu-fai Wong, "Nonlinear Scale-space Filtering and Multiresolution Systems," *IEEE Trans. Image Processing*, June, 1995.

[5] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. patt. Anal. Machine Intell.*, 13, 583–598, 1991.

[6] C. Gu and M. Kunt, "Contour simplification by a new nonlinear filter for region-based coding," *VCIP*, 1180-1191, 1994.

[7] S.G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. Patt. Anal. and Mach. Intell.*, 11, 674-693, 1989.

*This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

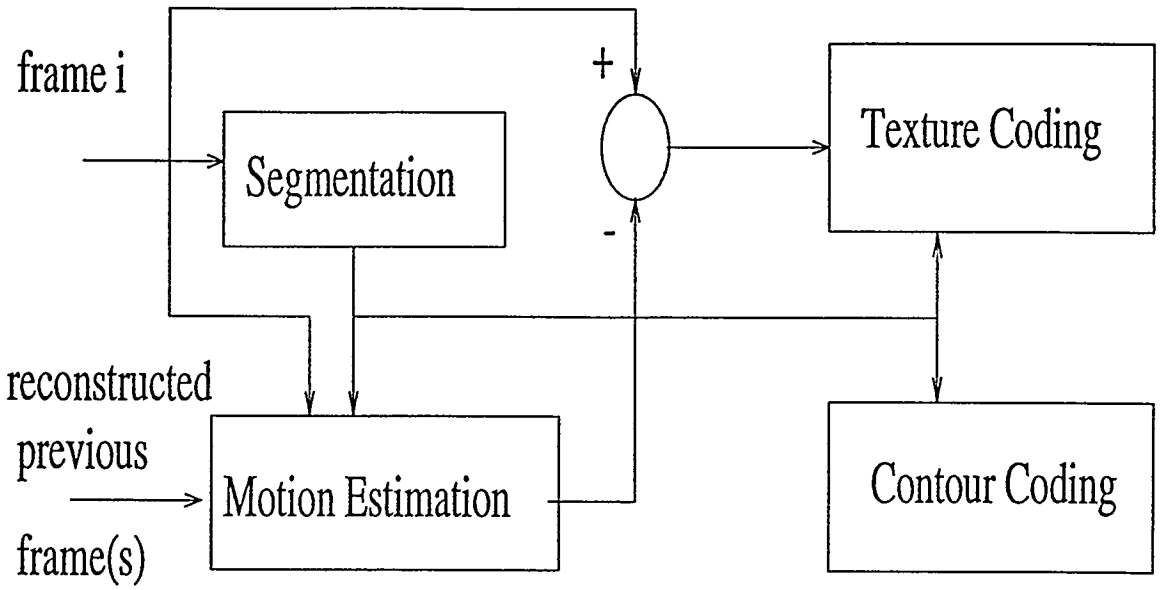


Figure 1. Schematic diagram of segmentation-based video coding.

Figure 2. compression ratio versus frame number for Miss America sequence

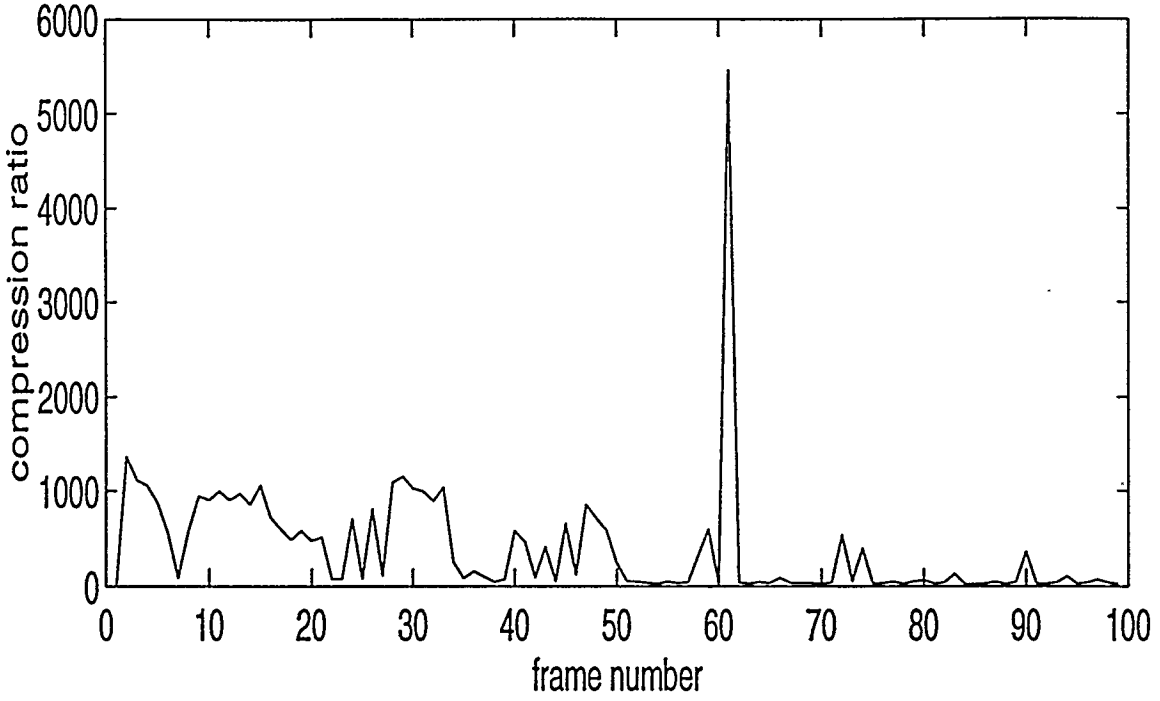


Figure 3. SNR versus frame number for Miss America sequence