Advanced Telecommunications and Signal Processing

Academic and Research Staff
Professor Jae S. Lim

Research Affiliate
Carlos Kennedy

Graduate Students
Zhenya Gu
Fatih Kamisli
Andy Lin
Han Wang

Support Staff
Cindy LeBlanc

Introduction

The present television system was designed nearly 60 years ago. Since then, there have been significant developments in technology, which are highly relevant to the television industries. For example, advances in the very large scale integration (VLSI) technology and signal processing theories make it feasible to incorporate frame-store memory and sophisticated signal processing capabilities in a television receiver at a reasonable cost. To exploit this new technology in developing future television systems, the research areas of the program focused on a number of issues related to digital television design. As a result of this effort, significant advances have already been made and these advances have been included in the U.S. digital television standard. Specifically, the ATSP group represented MIT in MIT's participation in the Grand Alliance, which consisted of MIT, AT&T, Zenith Electronics Corporation, General Instrument Corporation, David Sarnoff Research Center, Philips Laboratories, and Thomson Consumer Electronics. The Grand Alliance digital television system served as the basis for the U.S. Digital Television (DTV) standard, which was formally adopted by the U.S. Federal Communications Commission in December 1996.

The digital TV system based on this standard has been deployed successfully. In 2006, digital television receiver sales exceeded analog television receivers in both number and dollar volume in the U.S. The analog TV service is scheduled to discontinue in the spring of 2009.

The standard imposes substantial constraints on the way the digital television signal is transmitted and received. The standard also leaves considerable room for future improvements through technological advances. Future research will focus on making these improvements. The digital television system is a major improvement over the analog television system. The next major improvement over the digital television system is likely to be in the introduction of 3-D television. We are currently planning future research in this area.

The specific research topics where we made some progress are as follows:

1. Transforms for Prediction Residuals in Video Coding

Sponsor: Advanced Telecommunications Research Program

Project Staff: Fatih Kamisli

To store or transmit a large amount of data involved in visual media, digital image and video compression technologies are utilized. For example, to transmit the HDTV signal in a bandwidth originally designed for analog television broadcast, compression by a factor of 70 is performed. Compression is achieved in part by exploiting temporal and spatial redundancies present in visual media. Temporal redundancy refers to the fact that there are often very small changes from frame to frame within a video sequence. Spatial redundancy, present in a single frame or an image, refers to the similarity or slow variations of picture elements in a small neighborhood.

Spatial redundancy in images is reduced by applying transforms on a small neighborhood of pixels. These transforms have the energy-compaction property, which means that a small number of transform coefficients are sufficient to capture the signal in that small neighborhood with adequate fidelity. Temporal redundancy is reduced by using motion-compensated prediction techniques. Typically, a frame is divided into small blocks, and each block is predicted from previously transmitted frames by searching them for a good match for that block. The difference between the prediction and the frame to be coded is often called the motion-compensation residual (MC-residual). In standard video encoders, the MC-residual is compressed in the same way as an image is compressed with the widely used JPEG image compression. Specifically, the same transform, the Discrete Cosine Transforms (DCT), is used. The MC-residual is intimately connected to images it has been obtained from. However, its spatial characteristics differ considerably from that of an image. This research focuses on developing transforms for the MC-residual, as well as other residuals encountered in video coding, such as the upsampling residual in scalable video coding and the disparity-compensation residual in multiview video coding.

The properties of the MC-residual have been studied by various researchers [1-3]. In [1], the auto-covariance of the MC-residual is modeled as a sum of a first-order Markov-process and independent white noise. This model reflects the relatively weaker correlation of the MC-residual compared to images in a simple and tractable way. In [2], the authors propose another compound model which fits the tails of the auto-covariance of MC-residuals better than the model in [1]. A more complicated analysis resulting in a more complicated model is provided in [3]. All these studies indicate that the statistical characteristics of the MC-residual have differences from the statistical characteristics of images. However, transforms accounting for these differences can often not be derived directly from such characterizations because most of these characterizations are rather complicated.

Recently, there has been a great deal of research on transforms that can take advantage of locally anisotropic features in images [4-8]. Conventionally, the 2-D DCT or the 2-D Discrete Wavelet Transform (DWT) is carried out as a separable transform by cascading two 1-D transforms in the vertical and horizontal directions. This scheme does not take advantage of the locally anisotropic features present in images because it favors horizontal or vertical features over others. For example, the 2-D DWT has vanishing moments only in the horizontal and vertical directions. In these more recent approaches, transforms adapt to local anisotropic features by performing the filtering along the direction where the image intensity variations are smaller. This is achieved by resampling the image intensities along such directions [6], by performing filtering and subsampling on oriented sublattices of the sampling grid [7], by directional lifting implementations of the wavelet transform [4], or by various other means. Even though most of the work is based on the wavelet transform, applications of similar ideas to DCT-based image compression have also been made [8]. However these ideas have not yet been applied to modeling and compressing the MC-residual.

In this research, our goal is to develop transforms for the MC-residual as well as other residuals encountered in video coding, such as the upsampling residual in scalable video coding or the disparity-compensation residual in multiview video coding. Using insights obtained from the research on direction-adaptive image transforms, we investigate how locally anisotropic features of images affect the MC-residual. We obtain an adaptive auto-covariance characterization of the MC-residual, which reveals some statistical differences between the MC-residual and the image. Based on this characterization, we have developed a set of block transforms that can be used to compress the MC-residual. Future research will focus on obtaining similar characterizations and transforms for the upsampling and disparity estimation residuals.

References

[1] C.-F. Chen and K.K. Pang. The optimal transform of motion-compensated frame difference images in a hybrid coder. Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on [see also Circuits and Systems II:Express Briefs, IEEE Transactions on], 40(6):393–397, Jun 1993.

[2] W. Niehsen and M. Brunig. Covariance analysis of motion-compensated frame differences. Circuits and Systems for Video Technology, IEEE Transactions on, 9(4):536–539, Jun 1999.

[3] K.-C. Hui and W.-C. Siu. Extended analysis of motion-compensated frame difference for block-based motion prediction error. Image Processing, IEEE Transactions on, 16(5):1232–1245, May 2007.

[4] C.-L. Chang and B. Girod. Direction-adaptive discrete wavelet transform for image compression. Image Processing, IEEE Transactions on, 16(5):1289–1302, May 2007.

[5] E. Le Pennec and S. Mallat. Sparse geometric image representations with bandelets. Image Processing, IEEE Transactions on, 14(4):423–438, April 2005.

[6] D. Taubman and A. Zakhor. Orientation adaptive subband coding of images. Image Processing, IEEE Transactions on, 3(4):421–437, Jul 1994.

[7] V. Velisavljevic, B. Beferull-Lozano, M. Vetterli, and P.L. Dragotti. Directionlets: anisotropic multidirectional representation with separable filtering. Image Processing, IEEE Transactions on, 15(7):1916–1933, July 2006.

[8] Bing Zeng and Jingjing Fu. Directional discrete cosine transforms for image coding. Multimedia and Expo, 2006 IEEE International Conference on, pages 721–724, 9-12 July 2006.

2. Image Fusion: Increase of Depth-of-field by Combination of Multiple Images

Sponsor: Advanced Telecommunications Research Program

Project Staff: Andy Lin

Image fusion involves combining multiple images of one scene, where each image is obtained with a slightly different focal point. The objective is to produce a final image with a large depth-of-field and a high signal-to-noise ratio. This computational photography technique is useful in increasing the depth-of-field, without losing signal-to-noise ratio in low-light and macro/micro photography.

In traditional photography, image quality is subject to trade-offs under low-light situations. If a large depth-of-field is desired, the aperture must be closed substantially, resulting in a longer shutter-speed. Longer shutter-speeds often result in motion blur, so the sensor International Organization for Standardization (ISO) must be increased. This results in a noisier image. If a high signal-to-noise ratio is desired, the ISO must be set to the minimum. A low ISO requires more light, so the aperture must be opened, resulting in a small depth-of-field. Image fusion seeks to allow for a high signal-to-noise ratio, as well as a large depth-of-field.

In macro and micro photography, a large depth-of-field image is traditionally impossible, even if the aperture of the camera is almost completely closed. Image fusion seeks to produce high depth-of-field images in macro and micro photography.

Existing algorithms for image fusion fall into two categories, those that operate in the spatial domain and those that operate in the frequency domain. In both of these types of algorithms, a comparison is performed among the multiple images to select which image is sharpest at every pixel. In our research, we made various improvements to these algorithms.

One method involves using variable window sizes. Edge-detection is first performed on an initial decision map based on a large window size. Based on the distance from edges, a new window size is obtained at each pixel. Another method involves performing median-filtering on the decision map. Both techniques successfully reduce the mean square error of the resulting image by reducing noise in areas far from decision boundaries, without losing preciseness near decision boundaries.

This image fusion technique has potential applications in three-dimensional high-definition television. A potential approach for three-dimensional high-definition television is through depth-image-based rendering (DIBR), which requires a two-dimensional image and a per-pixel depth-map. Depth-image-based rendering allows for one way to incorporate multiple views and alternate camera-positions for three-dimensional high-definition television.

References

[1] Gabarda, Salvador, Gabriel Cristóbal, "On the use of a joint spatial-frequency representation for the fusion of multi-focus images," *Pattern Recognition Letters* **26(16):** 2572-2578 (2005).

[2] Levin, A. , Fergus, R., Durand, F., Freeman, B., "Image and Depth from a Conventional Camera with a Coded Aperture," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2007.

[3] Ng, Ren, M. Levoy, M. Bredif, G. Duval, M. Horowitz and P. Hanrahan, "Light Field Photography with a Hand-Held Plenoptic Camera," *Stanford University Computer Science Tech Report CSTR* 2005-02.

[4] Valdecasas, A.G., Marshall, D., Becerra, J.M., Terrero, J.J., 2001, "On the extended depth of focus algorithms for bright field microscopy," *Micron* **32**, 559-569.

[5] Z. Zhang and R. S. Blum, " A Hybrid Image Registration Technique for a Digital Camera Image Fusion Application," *Information Fusion*, pp. 135-149, June 2001.

3. Improvement of H.264 using JPEG 2000 Technology

Sponsor
Advanced Telecommunications Research Program

Project Staff
Zhenya Gu

JPEG is a commonly used still image compression standard developed in the late 1980s. It is widely used today for storing and transmitting digital images via the web. In the frequency domain, the image coefficients can be more efficiently compressed than in the spatial domain. JPEG uses a Fourier-related transform called the Discrete Cosine Transform (DCT) to convert images from the spatial to the frequency domain. Most of the loss in JPEG compression occurs during the quantization of the DCT coefficients. The coefficients are then entropy encoded for better compression.

A recent improvement on the JPEG standard is JPEG 2000. This new standard implements several new techniques to improve compression and image quality. One change is in the Fourier-related transform. JPEG 2000 uses the Discrete Wavelet Transform (DWT) instead of the DCT [1]-[2]. The DWT reduces blocking artifacts at low bitrates. However, the DWT has more blurring artifacts. JPEG 2000 also uses a binary arithmetic encoding for entropy encoding. This allows a noninteger number of bits for encoding each coefficient, as opposed to Huffman encoding, used by JPEG, which requires an integer number of bits for each coefficient.

In our research, we are studying how to apply the techniques of JPEG 2000 to H.264/MPEG-4, a video encoding standard. In video compression, there are three types of frames: I, P, and B frames. P and B frames are encoded using motion compensation, a technique in which previously encoded frames are used to predict the current frame. The error between the actual and predicted frame is encoded, not the frame itself. The I frames are coded independent of other frames. This creates a reference point, so that errors are not propagated through the entire video sequence.

H.264 uses more sophisticated motion compensation than older video encoding standards and context-adaptive binary arithmetic encoding. However, applying the DWT to H.264 may improve compression and image quality. In our research, we will apply many of the compression techniques used in JPEG2000 to the I frame of H.264 in order to improve compression and reduce low bitrate artifacts, such as blocking artifacts and color distortion. The two approaches can be compared using objective metrics, such as mean square error, or subjective metrics.

References

[1] M. Antonini, M. Barlaud, P.Mathieu, and I. Daubechies, "Image coding using the wavelet transform," *IEEE Trans. Image Processing* 1: 205-220 (1992).

[2] D. J. Le Gall and A. Tabatabai, "Subband coding of digital images using symmetric short kernel filters and arithmetic coding techniques," in *Proc. IEEE Int. Conf. ASSP*, NY, 761-765 (1988).

4. Transforms for Prediction Residuals in Video Coding

Sponsor: Advanced Telecommunications Research Program

Project Staff: Han Wang

Multi-view coding algorithms exploit the redundancy of information between subsequent image frames and also between the different views of the scene. The subjective quality of the reconstructed video sequences produced by such algorithms should be sufficiently high. The major compression techniques currently used for stereoscopic and multiview video coding include disparity compensation, 4-D sub-band coding [3], global geometric warping [4] and object-based schemes, which are for such specific applications as 3-D teleconferencing.

A major objective of multiview video compression/light field compression is to fully exploit the *intra-view* and *inter-view* coherence in the data set. The intra-view refers to the relationship among pixels within the same view, and inter-view refers to the relationship between pixels in views captured from different viewpoints. In addition, it is desirable to have a scalable representation of the video, which allows the system to efficiently adapt to varying storage capacities, transmission bandwidths, display devices, and computational resources by decompressing and rendering the video only up to a certain resolution, quality, or bit-rate requirement.

For two-view stereoscopic compression, the first left eye frame is often coded as an I-frame. Subsequent frames of the left-eye stream are predicted from the first frame using motion compensation and the right-eye frames are predicted from their corresponding left-eye frames using disparity compensation. Hence, the first frame of the left-eye stream is an I-frame and some additional frames are P-frames. In more sophisticated schemes, bi-direction predictions [1], similar to that in monocular streams, are often used. In light field compression, more complicated prediction schemes are exploited [2].

The global disparity is one of the major properties of the multiview video. The view disparity is similar to the temporal motion in the sense that they both represent the displacements between adjacent frames, whereas the properties and the inherent motion model are different. For example, it is well known that the disparity that represents the difference between two adjacent views is usually very compact. Therefore, a warping-based lifting transform can be used based on the assumption that the disparity between views can be well represented by a global disparity model. Since this assumption is not always true for the natural video due to the distortion of cameras and the scene depth, the practical solution should also consider the local disparity model.

Object-based coding of stereo sequences is an alternative to block-based schemes, since it can potentially produce fewer coding artifacts and provides a structural description of the scene useful in many applications. The block-based approach has the advantage of simplicity and robustness, allowing more straight-forward hardware implementations, but the subjective quality of reconstructed images may not be good in low bit rates. Comparing to the block-based approach, object-based schemes can alleviate the problem of annoying coding errors, providing a more natural representation of the scene, but require a complex analysis phase to segment the scene into objects and estimate their motion and structure.

Furthermore, important image areas such as facial details in face-to-face communications can be reconstructed with a higher image quality than with block-oriented hybrid coding. In addition, the ability of object-based coding techniques can describe a scene in a structural way, in contrast to traditional waveform-based coding techniques. Object-based approaches applied in coding of stereo image sequences have the additional benefit of conveying depth information which may be computed directly from the images. Using the depth information the scene may be separated in layers and depth keying is possible. Also accurate 3D modeling of the scene structure may be achieved. However, modeling techniques proposed in the literature are often restricted to video-phone sequences where the scene structure is known a priori and knowledge based parameterized models of face, arms, and body may be exploited. A more general approach exploits depth information.

In general, it is easy to see that each type of compression algorithm works well with classes of video frames with specific properties. In our research, we have developed an approach to combine various existing compression algorithms. Each algorithm works well with a relatively small class of images but with very good performance. The different algorithms are applied to each image and the algorithm that works best is selected for that image. This approach has a number of potential advantages over conventional methods where the same compensation and compression algorithm is used for all the frames in the video.

References

[1] Markus Flierl and Bernd Girod, “A New Bidirectional Motion-Compensated Orthogonal Transform for Video Coding,” pp. 665-668 *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, April, 2007.

[2] Marcus Magnor and Bernd Girod, “Data Compression for Light-Field Rendering,” *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 10, No. 3, April 2000.

[3] W. Yang, Y. Lu, F. Wu, J. Cai, K. N. Ngan and S. Li, “4-D Wavelet-Based Multiview Video Coding,” *IEEE Transactions On Circuits and Systems for Video Technology*, 16(11):1385-1396, November, 2006.

[4] F. Dufaux and J. Konrad, “Efficient, Robust, and Fast Global Motion Estimation for Video Coding,” *IEEE Transactions on Image Processing*, Vol. 9, No. 3, pp 497-501, March 2000.