Efficient Video Coding Considering a Video as 3D Data Cube

Manoranjan Paul
School of Computing and Mathematics
Charles Sturt University
Bathurst, Australia
Email: mmpaul@csu.edu.au

Weisi Lin
School of Computer Engineering
Nanyang Technological University
Singapore.
Email: wslin@ntu.edu.sg

Abstract—All existing video coding standards consider a video as a temporal (along T-axis) collection of two dimensional pictures (formed by XY axes) and compress them by exploiting spatial and temporal redundancy in the pictures. A recent optimal compression plane (OCP) determination technique shows that better compression can be achieved by relaxing the physical meaning of axes by exploring information redundancy in a fuller extent where a video is considered as a 3D data cube. Spatial and temporal dimensions are determined based on the statistical redundancy along each axis. Treating a video as a 3D data cube revolutionizes the traditional video features such as background, motion, object, zooming, panning, etc. In this paper we apply dynamic background modeling to the OCP plane to exploit the newly introduced background in the video for further improving the coding performance. The experimental results reveal that the proposed approach outperforms the existing state-of-the-art OCP technique as well as the H.264 video coding standard.

Keywords— Video coding, Background modelling, McFIS, Optimal compression plane, video preprocessing, and compression.

I. INTRODUCTION

Unlike any other multimedia component, a video requires a huge amount of raw data for storage and transmission to retain a pleasant quality. Thus, a proper compression technique (i.e., video coding) is inevitable to compress the huge amount of video data for practical applications in the stringent storage or bandwidth capacity. All existing video coding standards such as H.264 [1][2][3], MPEG-4 [4], etc. consider a video as a sequence of natural frames (formed by X and Y axes) and compress the frames by exploiting spatial redundancy (along X and Y axes within a frame) and temporal redundancy (along T axis among the frames) separately. Recently an optimal compression plane (OCP) determination technique [5] considers a video as a 3D data cube by the physical meaning of X, Y, and T-axes of a video and achieves more compression through the exploration of the information redundancy in an extensive way.

A video \( V_{XYT} \) can be represented with X, Y, and T-axes by considering any two axes to form a spatial frame and the rest axis to form temporal axis. Thus, a video can be represented as a collection of images \( I_{XY} \), \( I_{TX} \), or \( I_{TY} \) formed by XY, TX, or TY plane with T, Y, X axis as a temporal axis respectively:

\[
V_{XYT} = \{I_{XY}(t), t = 1,2,3,...\} \text{ or } V_{XYT} = \{I_{TX}(y), y = 1,2,3,...\} \text{ or } V_{XYT} = \{I_{TY}(x), x = 1,2,3,...\}
\]

where \( I_{XY} \) represents a natural image with axes of X and Y; \( I_{TX} \) represents a rearranged image with axes of T and X; \( I_{TY} \) represents a rearranged image with axes of T and Y. One example is given in Figure 1 where Figure 1(a) shows natural images of the Mobile video sequence. Figure 1(b) shows images using TX plane, and Figure 1(c) shows images using TY plane. The OCP technique [5] first determines the statistical redundancy along an axis using average correlation coefficients (CCs) [6] between frames formed by the two remainder axes. Then the OCP selects the optimal compression plane formed by the two axes which provide the largest CCs for intra-frame coding. A modified criterion for selecting the optimal compression plane is also proposed in [5] for inter-frame coding. The experimental results reveal that TX or TY plane rather than conventional XY plane is selected as an optimal compression plane for a large number of QCIF standard video sequences including as Mobile, Foreman, Container, Tempete, etc. An improved rate-distortion performance is achieved when JPEG2000 or H.264 coding technique is applied to the OCP.

Figure 1: Frames in XY and non-XY planes of Mobile video sequence [5].
Treating a video as a 3D data tube and rearranging the video frames in other two directions revolutionizes the traditional video features such as background, motion, object, panning, zooming, etc. Due to the rearrangement of the traditional XY plane images into the TX or TY plane images, object and/or camera motions of a traditional video can be transformed into a simplified motions or simple background, e.g., horizontal motions and vertical motions can be transformed into a static background in the TX or TY plane images respectively or a heterogeneous object can be transformed into a smooth object in TX or TY plane. Besides this, camera motions such as zooming and panning can be effectively estimated by the traditional translational motion estimation adopted into the H.264. Camera motions can also be transformed into a simplified motion/background in the TX or TY plane, thus any existing coding technique can encode them more efficiently.

Recently a dynamic background frame termed as the McFIS (the most common frame of a scene) [7] has been developed for video coding using the dynamic background modeling based on Gaussian mixture [8][9]. The McFIS is used as a second reference frame for encoding the current frame assuming that the motion part of the current frame would be referenced using the immediate previous frame and the static background part would be referenced using the McFIS. The ultimate reference is selected at block and sub-block level using the Lagrangian multiplier. The McFIS is used as a long term reference frame in the dual reference frames concept which is a subset of the concept of multiple reference frames. In this paper we like to exploit the newly introduced background in the TX or TY plane through the McFIS.

II. PROPOSED TECHNIQUE

The proposed technique has two phases (i) preprocessing phase to determine the optimal compression plane and (ii) actual coding phase to encode the frame with the help of McFIS modeling. In the preprocessing phase, the proposed technique determines the coding direction of a video using the statistical redundancy. The coding direction would be one of three possible directions such as $V_{XY}$, $V_{TY}$, or $V_{TX}$ where first two axes form the spatial image and the third axis forms the temporal image. In the actual coding, a frame is encoded using two reference frames; one is the immediate previous frame and the other one is the McFIS assuming that object areas and the background areas of the current frame use the immediate previous frame and the McFIS as the reference frame respectively based on the Lagrangian multiplier. The McFIS (i.e., a dynamic background frame) is generated using the encoded frames at the encoder and the decoder. As the same procedure and the same encoded frames (assuming that no error is occurred during the channel coding) are used at the encoder and the decoder, we do not need to transmit the McFIS from encoder to the decoder. The k-th McFIS is used to encode the (k+1)-th frame where k number of encoded frames are used to form the k-th McFIS based on the Gaussian Mixture modeling. The OCP formation and the McFIS generation is detailed at the next two subsections.

A. Optimal Compression Plane

As mentioned before a video can be treated as a 3D data cube and represented as one of $V_{XY}$, $V_{TY}$, or $V_{TX}$ formats. Obviously the most accurate determination of compression plane of a video is to apply the H.264 on each of the formats and then select the best one based on the joint rate-distortion performance. This brute-force method takes three times computational time compared to the standard coding schemes. To reduce the computational time, an OCP determination algorithm [5] is proposed based on the statistical redundancy. The statistical redundancy along one axis can be estimated by the average CC formed by the other two axes. The bigger the average CC is, the more the statistical redundancy exists [5]. The CC is defined in [6] as follows:

$$\text{CC}_{k,t+1} = \frac{\sum_{i,j} (I_k(x,y) - \overline{I_k}) \sum_{i,j} (I_{k+1}(x,y) - \overline{I_{k+1}})}{\sqrt{\sum_{i,j} (I_k(x,y) - \overline{I_k})^2 \sum_{i,j} (I_{k+1}(x,y) - \overline{I_{k+1}})^2}}$$

where, $I_k(x,y)$ and $I_{k+1}(x,y)$ are the pixel intensities located at the (x,y) position of k-th and (k+1)-th frames respectively and $\overline{I_k}$ and $\overline{I_{k+1}}$ are the average pixel intensities of k- and (k+1)-th frames respectively.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

B. McFIS Generation

We assume that k-th Gaussian at time t representing a pixel intensity with mean $\mu_t^k$, standard deviation (STD) $\sigma_t^k$, recent value $\gamma_t^k$, and weight $\alpha_t^k$ such that $\sum \alpha_t^k = 1$. The learning parameter $\alpha$ is used to balance the contribution between the current and past values of parameters such as weight, STD, mean, etc. After initialization, for every new observation $X_t$ (pixel intensity at time t) is first matched against the existing models in order to find one (e.g., k-th model) such that $|X_t - \mu_{k-1}^k| \leq 2.5 \sigma_{k-1}^k$. If such a model exists, update corresponding recent value parameter $\gamma_{k,t}$ with $X_t$. Other parameters are updated with learning rate as:

$$\mu_{t}^k = (1-\alpha) \mu_{t-1}^k + \alpha X_t$$
$$\sigma_{t}^2 = (1-\alpha) \sigma_{t-1}^2 + \alpha (X_t - \mu_{t}^k)^2 (X_{t-1} - \mu_{t-1}^k)$$
$$\alpha_{t}^k = (1-\alpha) \alpha_{t-1}^k + \alpha$$

and the weights of the remaining Gaussians (i.e., l where $l \neq k$) are updated as $\alpha_l^t = (1-\alpha) \alpha_{t-1}^l$. Afterward, the weights are renormalized. If such a model does not exist, a new Gaussian model is introduced with $\gamma = X_t$, $\sigma = 30$, and $\omega = 0.001$ by evicting the K-th (based on wfr in descending order) model if it exists. For detailed procedure, please refer.

To get the background pixel intensity from the above mentioned DBM technique for a particular pixel, we take average of the mean pixel value and recent pixel value of the model that has the highest value of weight/standard deviation.
a background frame (comprising background pixels) as the McFIS. Two examples of McFIS are shown in Error! Reference source not found., using first 16 original frames of Paris and Silent video sequences respectively. Error! Reference source not found. (a) & (b) show the 16th frames of corresponding videos and (c) & (d) show McFISes. The circles in (c) & (d) indicate the uncovered/occluded background captured by the corresponding McFIS. To capture the uncovered background by any single frame is quite impossible unless this uncovered background is visible for one frame and that frame is used as LTR frame in the relevant existing approaches. A McFIS is more suitable as a LTR frame than any single pre-encoded frame.

III. PROPOSED TECHNIQUE

A. McFISes in XY, TX, and TY Planes

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

IV. EXPERIMENTAL RESULTS

To compare the performance of the proposed HBP scheme using McFIS as a third reference frame, we have also implemented the original HBP scheme, the HBP scheme with an extra reference frame (HBP-3Ref), and the HBP scheme with LTR frame (HBP-LTR). All the algorithms are implemented based on the H.264 recommendations with 25 Hz, ±15 as the search length with quarter-pel accuracy, as 16 as GOP size. The proposed technique has three reference frames and the original HBP technique has two reference frames. Thus, for fair comparison we have selected HBP-3Ref scheme. We also like to see the effectiveness of the McFIS over the LTR frame and the first frame of a scene as a third reference frame in the HBP scheme. To verify this we have selected the HBP-LTR (20 jumping parameter) and HBP-FirstFrame (i.e., LTR frame is set to the first frame of a scene) schemes. We have used same QP for high quality LTR frame and the McFIS for fair comparison. We use first 288 frames for videos except the high resolution (4CIF) Popple video (with 112 frames).

During the preprocessing phase we need to process a number of frames before the actual coding, he proposed technique has limited application especially when the system requires real time duplex communication.

Obviously the proposed technique will take some extra operations to generate and encode McFIS compared to the conventional HBP scheme with three reference frames. As we generate and encode the McFIS once for a scene using first few frames, the impact of extra operations is negligible (only 1%) more compared to the H.264-3Ref scheme as shown in Error! Reference source not found.. Error! Reference source not found. shows 16~50% and 3~28% increase of skip MBs using the proposed and the HBP-LTR schemes compared to the HBP-3Ref scheme as the McFIS represents the stable part of a scene (i.e., background). This figure indicates that the more reduction of bit rates by the proposed technique compared to the others.

A. Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

1) For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.
   a) Selection (Heading 4): Highlight all author and affiliation lines.

   b) Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

   c) Deletion: Delete the author and affiliation lines for the second affiliation.

   d) For author/s of more than two affiliations: To change the default, adjust the template as follows.

   e) Selection: Highlight all author and affiliation lines.

   f) Change number of columns: Select the “Columns” icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

   g) Highlight author and affiliation lines of affiliation 1 and copy this selection.

   h) Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste down the copy of affiliation 1. Repeat as necessary for each additional affiliation.

   i) Reassign number of columns: Place your cursor to the right of the last character of the last affiliation line of an even numbered affiliation (e.g., if there are five affiliations, place your cursor at end of fourth affiliation). Drag the cursor up to highlight all of the above author and affiliation lines. Go to Column icon and select “2 Columns”. If you have an odd number of affiliations, the final affiliation will be centered on the page; all previous will be in two columns.

B. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head
because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

C. Figures and Tables

1) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert with a ratio of quantities and units. For example, write abbreviation “Fig. 1”, even at the beginning of a sentence. Across both columns. Figure captions should be below “Temperature (K)”, not “Temperature/K”.

Table Labels: Use 8 point Times New Roman for Table labels. In an MSW document, this method is somewhat more stable than directly inserting a graphic. We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization,” or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization [A[m(1)])],” not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

ACKNOWLEDGMENT (HEADING 5)

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R. B. G.) thanks . . . .” Instead, try “R. G. thanks”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

Figure 2. Rate-distortion performance by the proposed, the optimal compression plane, and the H.264 techniques using six standard video sequences.