HUMAN VISION SYSTEM AWARE EXHAUSTIVE BLOCK-MATCHING ALGORITHM

David Chih-Che Lin and Paul M. Chau

University of California, San Diego
Department of Electrical and Computer Engineering
9500 Gilman Drive, Mail Code 0407
La Jolla, CA 92093-0407

ABSTRACT

In this work, homomorphic image modeling is used to make the exhaustive block-matching algorithm (EBMA) more human vision system (HVS) aware, thus yielding visually pleasing sequences. Homomorphic signal processing is used to separate the luminous and the structural components of each frame; EBMA is then applied to both components to capture the luminous and the structural changes. The combination of these two techniques is utilized to simulate the structure preserving nature of HVS. Temporal irrelevancies are reduced by removing the excess motion fields from both components. The homomorphic EBMA (H-EBMA) processed sequences have superior visual quality despite mean PSNR values that are comparable to sequences produced by traditional EBMA.

1. INTRODUCTION

In most video compression systems, temporal redundancies between transmitted frames are removed by subtracting a predicted frame from the current frame. The accuracy of the predicted frame is improved through estimating the motion between the reference and current frame. The residual frame and the motion vectors are encoded and transmitted to the decoder. The decoder then recreates the prediction frame (with the motion vectors) and combines it with the difference frame to reconstruct the current frame. A better motion estimation algorithm would greatly reduce the energy in the residual frame, thereby, increasing the overall compression ratio.

An important component of video compression is to find an efficient way to estimate motion. For natural video sequences consisting of multiple objects, the motion field can be optimally described by assigning each object its own motion vector. Due to the complexity of such region-based motion estimation algorithms, practical video compression systems (such as H.26x) choose to implement the translational block-based representation instead [6]. In this motion representation, the frame is divided into numerous sub-blocks and each sub-block is assumed to undergo translational motion only. The exhaustive block-matching algorithm (EBMA) is the simplest form of this model. Consequentially, most video compression systems implement some variation of EBMA.

While EBMA based algorithms adequately describe motion fields and reduce temporal redundancies, more compression can be achieved if temporal irrelevancies are also exploited. To properly remove these irrelevancies, the compression systems need to include a human visual system (HVS) model. Implementation of such models in real-time video compression systems are not feasible due to the complexity and non-linearity of most HVS models. In this work, the EBMA is made more HVS aware by incorporating a simple hypothesis about the HVS proposed in [5]: *The main objective of human visual system is to perceive structures of objects.* To include the hypothesis, the algorithm uses the homomorphic image model presented in [2] to isolate the structure and luminance of the current frame. EBMA is subsequently applied to both luminous and structural sub-frames. The decoder recreates both sub-frames and combines the two to reconstruct the current frame. The homomorphic EBMA (H-EBMA), proposed in this work, produces video sequences of superior visual quality while mean PSNR values are comparable to sequences processed by EBMA alone. Furthermore, temporal irrelevancies are reduced by eliminating redundant luminious/structural motion fields.

This paper is organized as follows: Part 2 highlights the EBMA. Part 3 overviews the homomorphic image model. Part 4 describes the proposed H-EBMA. Part 5 contains experimental results and Part 6 concludes the contributions of this work.

2. EBMA

EBMA is designed to determine a matching block, B_R, in the reference frame given the block, B_C, in the current frame such that the error between these two blocks are minimized [6]. The displacement between these two blocks, $\mathbf{d_j}$, is the motion vector of B_C. If a frame in the video sequence is divided into $\mathfrak{J}$ blocks, then the over all error can be described as:

$$e(\mathbf{d_j}, \forall j \in \mathfrak{J}) = \sum_{j \in \mathfrak{J}} \sum_{\mathbf{m} \in B_C} \mid x_R(\mathbf{m} + \mathbf{d_j}) - x_C(\mathbf{m}) \mid^p \quad (1)$$

Since the estimated motion vector only affects the prediction error of that block only, the motion vector for each block can be estimated independent of the other blocks. Therefore, equation (1) can also be written as:

$$\begin{aligned} e(\mathbf{d_j}, \forall j \in \mathfrak{J}) &= \mathfrak{J} \cdot e_j(\mathbf{d_j}) \\ &= \mathfrak{J} \cdot \sum_{\mathbf{m} \in B_C} \mid x_R(\mathbf{m} + \mathbf{d_j}) - x_C(\mathbf{m}) \mid^p \end{aligned}$$

In EBMA, the current block, B_C is compared against all possible blocks in the designated search range of the reference frame, $\forall B_R \in \mathbf{S}$, to obtain the $\mathbf{d_j}$ that minimizes the block error in equation (1). To reduce the complexity of calculation, the minimum absolute difference (MAD) error, $p = 1$ in equation (1), is used in this work.

Each iteration to compute the MAD requires three computations: absolute value, subtraction and addition. In [6], it is shown that each MAD computation requires approximately $3N^2(2S+1)^2$ computation steps. $3M^2(2S+1)^2$ computation steps are required to perform EBMA for the whole image. (Assuming block size of NxN pixels, search range of SxS pixels and the image size of MxM pixels where M is a multiple of N.)

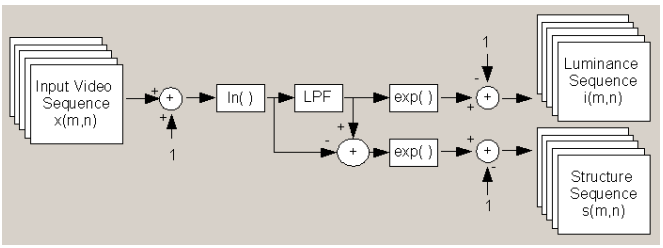

Fig. 1. Pre-processing

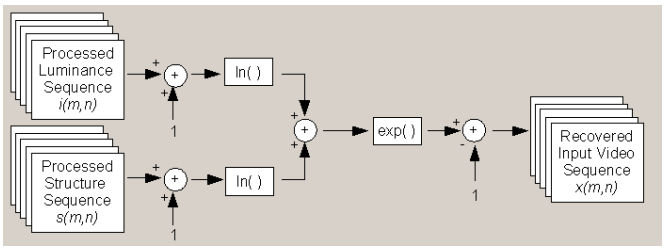

Fig. 2. Post-processing

3. HOMOMORPHIC IMAGE MODEL

Generally images are produced from some physical processes. The two most important processes are the source illumination and object reflection. The illumination is usually generated from an electromagnetic energy source (radar, infrared, natural lights, ... etc.) The illumination is either transmitted through or reflected by objects. In this sense, any image, $x(\mathbf{m})$, can be viewed as a product of two sub-images:

$$x(\mathbf{m}) = i(\mathbf{m}) \cdot s(\mathbf{m}) \quad (2)$$

$i(\mathbf{m})$ refers to the illumination (a property of the illuminating source) and $s(\mathbf{m})$ refers to the reflectance of objects (which is directly related to the structures of objects). These two components can be separated by homomorphic signal processing. If we define $\hat{x}(\mathbf{m}) = ln[x(\mathbf{m})]$ then:

$$\begin{aligned} \mathfrak{F}\{\hat{x}(\mathbf{m})\} &= \mathfrak{F}\{ln(i(\mathbf{m}) \cdot s(\mathbf{m}))\} \\ &= \mathfrak{F}\{\hat{i}(\mathbf{m})\} + \mathfrak{F}\{\hat{s}(\mathbf{m})\} \\ \hat{X}(\mathbf{\Omega}) &= \hat{I}(\mathbf{\Omega}) + \hat{S}(\mathbf{\Omega}) \end{aligned} \quad (3)$$

The benefit of this model is that in the logarithmic space, the structural, $\hat{s}(\mathbf{m})$, and luminous, $\hat{i}(\mathbf{m})$, sub-images are separable in the frequency domain. The luminance, $\hat{i}(\mathbf{m})$, is the low frequency component and the structure, $\hat{s}(\mathbf{m})$, is the mid and high frequency components. Homomorphic image model has mainly been used in image restoration [2, 7]. Other attempts to incorporate the homomorphic image model into video processing can be found in [3, 4].

4. PROPOSED HVS AWARE EBMA

Each pixel in a grayscale image typically requires 8bpp, which translates to a dynamic range of 0 to 255. The previous works on homomorphic image modeling ([2, 3, 4]) all ignored the 0 during the transformation into the logarithmic space. In this work, an offset of one is used before transforming the frame into the logarithmic space. This offset of one is subsequently removed during the transformation back to the Cartesian number space. In the H-EBMA, the luminous and structural images are separated in the pre-processing stage in a similar fashion as the steps described in [4]. In the first section, ideal motion estimation is assumed and perfect reconstruction (PR) is achieved despite of the offset by one and the pre and post processing steps. The PR condition for imperfect motion estimation and proposed steps to remove temporal irrelevancies are outlined in the subsequent sections.

4.1. System Analysis with Perfect Motion Estimation

In the pre-processing step (figure 1), each frame of the video sequence, $x(\mathbf{m})$, is first transformed into the logarithmic space where the luminous and structural components are separated by a low pass filter, with impulse response of $h(\mathbf{m})$, as the following:

$$\begin{aligned} \hat{I}(\mathbf{\Omega}) &= H(\mathbf{\Omega})ln(X(\mathbf{\Omega}) + 1) \\ \hat{S}(\mathbf{\Omega}) &= (1 - H(\mathbf{\Omega}))ln(X(\mathbf{\Omega}) + 1) \end{aligned}$$

The two components are transformed back to the real space as the following:

$$\begin{aligned} I(\mathbf{\Omega}) &= exp(\hat{I}(\mathbf{\Omega})) - 1 \quad (4) \\ &= exp(H(\mathbf{\Omega})ln(X(\mathbf{\Omega}) + 1)) - 1 \\ S(\mathbf{\Omega}) &= exp(\hat{S}(\mathbf{\Omega})) - 1 \quad (5) \\ &= \frac{X(\mathbf{\Omega}) + 1}{exp(H(\mathbf{\Omega})ln(X(\mathbf{\Omega}) + 1))} - 1 \end{aligned}$$

Assuming the motion estimation process is lossless, the estimated luminous and structural components, $\tilde{I}(\mathbf{\Omega})$ and $\tilde{S}(\mathbf{\Omega})$ are equal to the original luminous and structural components. Therefore, PR is achieved and the reconstructed frame in the post-processing step (figure 2), $\tilde{X}(\mathbf{\Omega})$ is:

$$\begin{aligned} \hat{\tilde{I}}(\mathbf{\Omega}) &= H(\mathbf{\Omega})ln(X(\mathbf{\Omega}) + 1) \\ \hat{\tilde{S}}(\mathbf{\Omega}) &= (1 - H(\mathbf{\Omega}))ln(X(\mathbf{\Omega}) + 1) \\ \tilde{X}(\mathbf{\Omega}) &= exp(\hat{\tilde{I}}(\mathbf{\Omega}) + \hat{\tilde{S}}(\mathbf{\Omega})) - 1 \\ &= exp(ln(X(\mathbf{\Omega}) + 1)) - 1 \\ &= X(\mathbf{\Omega}) \end{aligned}$$

4.2. System Analysis with Imperfect Motion Estimation

With imperfect motion estimation, the estimated luminance and structure become:

$$\begin{aligned} \tilde{I}(\mathbf{\Omega}) &= I(\mathbf{\Omega}) + E_I(\mathbf{\Omega}) \quad (6) \\ &= exp(H(\mathbf{\Omega})ln(X(\mathbf{\Omega}) + 1)) - 1 + E_I(\mathbf{\Omega}) \\ \tilde{S}(\mathbf{\Omega}) &= S(\mathbf{\Omega}) + E_S(\mathbf{\Omega}) \quad (7) \\ &= \frac{X(\mathbf{\Omega}) + 1}{H(\mathbf{\Omega})ln(X(\mathbf{\Omega}) + 1)} - 1 + E_S(\mathbf{\Omega}) \end{aligned}$$

where $E_I(\mathbf{\Omega})$ and $E_S(\mathbf{\Omega})$ are the errors caused by imperfect motion estimation. The reconstructed image, $X(\mathbf{\Omega})$, in this case becomes:

$$\begin{aligned} \hat{\tilde{I}}(\mathbf{\Omega}) &= ln(\tilde{I}(\mathbf{\Omega}) + 1) \\ &= ln(exp(H(\mathbf{\Omega})ln(X(\mathbf{\Omega}) + 1)) + E_I(\mathbf{\Omega})) \\ \hat{\tilde{S}}(\mathbf{\Omega}) &= ln(\tilde{S}(\mathbf{\Omega}) + 1) \\ &= ln\left(\frac{X(\mathbf{\Omega}) + 1}{H(\mathbf{\Omega})ln(X(\mathbf{\Omega}) + 1)} + E_S(\mathbf{\Omega})\right) \end{aligned}$$

$$
\begin{aligned}
X(\mathbf{\Omega}) &= exp(\hat{\tilde{I}}(\mathbf{\Omega}) + \hat{\tilde{S}}(\mathbf{\Omega})) - 1 \\
&= (exp(H(\mathbf{\Omega})ln(X(\mathbf{\Omega})+1))exp(E_I(\mathbf{\Omega}))) \\
&\quad \cdot \left(\frac{exp(ln(X(\mathbf{\Omega})+1))exp(E_S(\mathbf{\Omega}))}{exp(H(\mathbf{\Omega})ln(X(\mathbf{\Omega})+1))} \right) - 1 \\
&= (X(\mathbf{\Omega})+1)exp(E_I(\mathbf{\Omega}) + E_S(\mathbf{\Omega})) - 1
\end{aligned} \tag{8}
$$

Perfect reconstruction would be possible if:

$$
\begin{aligned}
E_I(\mathbf{\Omega}) &\sim 0 \\
E_S(\mathbf{\Omega}) &\sim 0
\end{aligned}
$$

Since the error in one block does not affect the error in the other blocks, minimizing the above can be described as minimizing $E_{j_I}(\mathbf{\Omega})$ and $E_{j_S}(\mathbf{\Omega})$, which is the same as minimizing their inverse fourier transforms: $e_{j_I}(\mathbf{d_j})$ and $e_{j_S}(\mathbf{d_j})$.

$$
\begin{aligned}
e_{j_I} &= \sum_{\mathbf{m} \in B_M} \mid i_R(\mathbf{m}+\mathbf{d_j}) - i_C(\mathbf{m}) \mid \\
e_{j_I}^2 &= \sum_{\mathbf{m} \in B_M} \mid i_R(\mathbf{m}+\mathbf{d_j}) - i_C(\mathbf{m}) \mid^2
\end{aligned} \tag{9}
$$

From Parseval's Theorm, equation (9) can be written as:

$$
\begin{aligned}
e_{j_I}^2 &= \sum_{\mathbf{\Omega}} \mid I_R(\mathbf{\Omega})exp(j\mathbf{\Omega}^T\mathbf{d_j}) - I_C(\mathbf{\Omega}) \mid^2 \\
e_{j_I} &= \sum_{\mathbf{\Omega}} \mid I_R(\mathbf{\Omega})exp(j\mathbf{\Omega}^T\mathbf{d_j}) - I_C(\mathbf{\Omega}) \mid
\end{aligned}
$$

$I_C(\mathbf{\Omega}) \sim I_R(\mathbf{\Omega})$ as long as $\mathbf{d_j} \sim 0$, yielding $e_{j_I} \sim 0$, thus PR for the luminous plane is possible. Similar analysis can be performed for the structural plane. Therefore, the overall system is approximately PR.

4.3. Temporal Irrelevancies Reduction

In natural video sequences, changes in subsequent frames can be attributed to a combination of luminous or structural changes. In some sequences, changes in illumination cause the same object in subsequent frames to appear different [7]. This aspect of HVS is implemented in [3], where the authors assume that structural changes can be estimated from changes in illumination. However, the authors' assumption failed for the case of structural changes under constant illumination. In this work, structural changes are tracked along with the luminous changes to include the structure detecting essence of HVS [5].

In the case where the component change is insubstantial, temporal irrelevancies can be removed by ignoring the change and keeping that component constant. The procedures are described below:

1. The net luminous motion, $\mathbf{D_I}$, and the net structural motion, $\mathbf{D_S}$ are defined as:

$$
\begin{aligned}
\mathbf{D_I} &= \sum_{\forall j \in \mathfrak{J}} \mathbf{d_{j_I}} \\
\mathbf{D_S} &= \sum_{\forall j \in \mathfrak{J}} \mathbf{d_{j_S}}
\end{aligned}
$$

   For each frame, $\mathbf{D_I}$ and $\mathbf{D_S}$ are stored into buffers: $\mathbf{D_{I}}_B$ and $\mathbf{D_{S}}_B$.

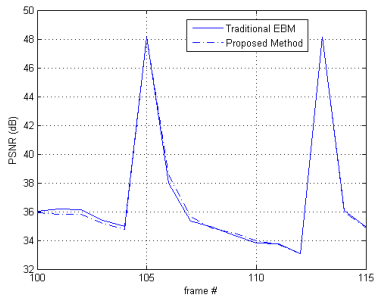

Fig. 3. EBMA vs. H-EBMA

2. For the next frame, luminous/structural motion is deemed insignificant if

$$
\begin{aligned}
\mid \mathbf{D_I} - \mathbf{D_{I}}_B \mid &< T \\
\mid \mathbf{D_S} - \mathbf{D_{S}}_B \mid &< T
\end{aligned}
$$

   where T is a preset threshold.

If the current luminous/structural motion is determined to be irrelevant, then the change is negligible, resulting in that sub frame(s) being replaced with the previous sub frame(s).

5. EXPERIMENTAL RESULTS

A series of experiments have been performed to evaluate the performance of the H-EBMA. All the experiments have been performed in MATLAB. In the experiments, every 8th frame was an I frame. In order to properly demonstrate the effectiveness of H-EBMA, the predicted frames were only reconstructed with the obtained motion fields. The analysis of the system demonstrated that the choice of the low pass filter has no effect on the overall performance. In these experiments, a 5x5 Gaussian smoothing filter (with $\sigma = 15$) was used to separate the luminous and structural components. The search range was set to 8x8 pixels. The simulation results for one of the test sequences, the 400 frames Foreman sequence of QCIF (176x144 pixels) resolution, are summarized in this section.

5.1. EBMA vs. H-EBMA

(a) Original (b) EBM (c) H-EBM

Fig. 4. Comparison between EBMA and H-EBMA. Search range of 8 x 8 pixels and block size of 4 x 4 pixels

This experiment was designed to compare the performance of the EBMA and the H-EBMA. In this experiment, both EBMA and H-EBMA were used to process the test video sequence. The search block size was set to 4x4 pixels in both algorithms. In this experiment, the threshold was set to zero for the H-EBMA as all luminous

Table 1. Effects of threshold

Threshold	MPSNR (dB)	% Luminance	% Structure
50	35.0	94.3	93.4
100	34.8	91.7	88.3
500	33.5	68.3	62.0
1000	32.1	50.0	45.2
1500	30.8	38.0	32.0
2000	30.1	31.7	28.3
5000	27.9	16.3	9.7

Table 2. Effect of various luminance and structure block size

L Block size	MPSNR (dB)	% Luminance	% Structure
2x2	33.2	81.4	45.1
4x4	32.1	50.0	45.1
S Block size	MPSNR (dB)	% Luminance	% Structure
2x2	34.2	50.0	66.3
4x4	32.1	50.0	45.1

and structural motion fields were used in the reconstruction of the predicted frames.

The resulting PSNR values for a section of the test sequence can be seen in figure 3. While sometimes one algorithm outperformed the other, both algorithms generally produced predicted frames of similar PSNR values. Zoomed in versions of the same resulting frame from both algorithms can be seen in figure 4. In this case, the H-EBMA predicted frame has a PSNR value of 26.4 dB and the EBMA predicted frame has a PSNR value of 26.0 dB. While the two frames have similar PSNR values, the H-EBMA predicted frame is more visually pleasing. Specifically, the eyes, the nose and the finger in front of the mouth were better preserved by the H-EBMA. The results show that the H-EBMA is a more HVS aware motion estimation algorithm.

5.2. Reduction of Temporal Irrelevancies

This experiment was designed to examine the effect of the varying the pre-set threshold on the quality of reconstructed video sequence in the H-EBMA. The various thresholds and the resulting mean PSNR (MPSNR) values are shown in table 1. The percentages of luminous and structural motion fields used in processing the test sequence are also listed in the table.

The results show that the increase of the pre-set threshold leads to a significant reduction in the retained information with a penalty in loss of MPSNR value. The optimal threshold for the test sequence was determined to be at around 500 through visual observation of the reconstructed sequences. When the threshold is set too high, the system becomes insensitive to changes. Consequently, too many motion fields are removed and a loss in the overall temporal smoothness of the video sequence is observed. However, when the threshold is set appropriately, H-EBMA successfully removes irrelevancies from the video sequence without a loss in perceptual quality. If the irrelevant frames are removed instead, the H-EBMA transforms into an efficient and HVS aware algorithm for frame rate down conversion.

5.3. Structural Preservation

Two separate experiments were designed to compare the importance of luminous and structural components in video compression. To better examine the effect of preserving one of the two components, the pre-set threshold was set to 1000. In the first experiment, the luminance block size was decreased to 2x2 pixels while the structure block size was kept constant at 4x4 pixels. The exact opposite was performed in the second experiment. The results from both experiments are shown in table 2.

From the results, it is seen that with roughly 50% of one component stored, the scheme that preserves structure yields higher quality video sequence. From the experiments, storing only 16% more of structure increases the MPSNR value by 2dB. On the contrary, a mere 1dB increase in MPSNR value is observed with 30% more of luminance stored. The structure preserved video sequence has also higher visual quality comparing to its counter part. The results suggest that structural preservation is far more important to the quality of the reconstructed video comparing to luminous preservation. This conclusion is consistent with the simplified hypothesis regarding the HVS system presented in [5].

6. CONCLUSION

In this work, the homomorphic image model, which allows for the separation of structure and luminance in the logarithmic space, is integrated into EBMA to create a more HVS aware motion estimation algorithm. The H-EBMA is HVS aware since it produces visually superior video sequences comparing to sequences processed by the EBMA. It is realized in this work that motion estimation in the logarithmic space yields motion vectors with actual physical meanings. These structural and luminous motion vectors are used in the proposed H-EBMA to remove the temporal irrelevancies from the video sequence. Lastly, in video compression systems, structural preservation leads to higher signal to noise ratio and better visual quality video sequences.

7. REFERENCES

[1] C. Steller, et. al, *Estimating Motion in Image Sequences*, IEEE Signal Processing Magazine (July 1999), 16:70-91

[2] A. Oppenheim, et. al, *Nonlinear Filtering of Multiplied and Convolved Signals*, Proceedings of the IEEE, 56(8):12641291, 1968

[3] H. Gomez-Moreno, et. al, *Extraction Illumination From Images Using the Wavelet Transform*, ICIP 2001, Oct. 2001, Greece

[4] T. Toth, et. al, *Illumination-Invariant Change Detection*, Proc. 4th IEEE Southwest Symp. Image Analysis and Interpretation, Apr. 2000, pp. 37.

[5] Z. Wang et. al, *Image Quality Assessment: From Error Visibility to Structural Similarity*, IEEE trans. on Image Processing, Vol 13, No. 4, April 2004

[6] Y. Wang, et. al, *Video Processing and Communications*, Prentice-Hall, 2001

[7] R.C. Gonzalez et. al, *Digital Image Processing*, Addison-Wesley, 1987

[8] I. Richardson, *H.264 and MPEG-4 Video Compression*, Wiley, 2003