AN EYE-MOVEMENT CONTROLLED WAVELET PACKET BASED IMAGE CODER

Peter Bergström

Image Coding Group, Dept. of Electrical Engineering, Linköping University, Sweden
peter@isy.liu.se

ABSTRACT

Abstract. An image coding scheme which combines transform coding with a human visual system (HVS) model has been developed. The system includes an eye tracker to pick up the point of regard of a single viewer. One can then utilize that the acuity of the HVS is lower in the peripheral vision than in the central part of the visual field. A model of the decreasing acuity of the HVS which can be applied to a wide class of transform coders is described. Such a coding system has large potential for data compression.

In this paper we have incorporated the model into an image coder based on the discrete wavelet packet transform (DWPT) scheme.

1. INTRODUCTION

The field of image coding deals with efficient ways of representing images for transmission and storage. Most image coding methods have been developed for TV-distribution, teleconferencing and video phones. Few efforts have been devoted towards coding methods for interactive systems. One example where interactive systems exists is in tele-robotics, where a human operator controls a robot at a distance. Interactive systems usually have only one observer of the transmitted image. In such a system one can include an eye tracker to pick up the point of regard of the viewer.

The human visual system (HVS) works as a space variant sensor system providing detailed information only in the gaze direction. The sensitivity decreases with increasing eccentricity and is much lower in the peripheral visual field. Thus, in a system with a single observer whose point of gaze is known, one can allow the image to be coded with decreasing quality towards the peripheral visual field.

In previous works we have incorporated the model of the HVS’s acuity described in section 2 into the JPEG coder with good results [1] and a wavelet based coder [2] with not as good results as for the modified JPEG coder. In this work we will apply the HVS model to a coder based on the discrete wavelet packet transform (DWPT) decomposition scheme.

The main steps in the transform coders we have used are transformation, scalar quantization, scanning and finally entropy coding. In the JPEG coder the image is split into 8x8 blocks. Each block is discrete cosine transformed into 64 transform component [3]. In the wavelet based coder the image is decomposed into an octave-band representation [4]. A DWPT decomposition scheme offers an adaptive wavelet decomposition. If a subband will be divided further or not is decided on a cost function. Thus, the resulting decomposition will depend on the image, the decomposition cost function and the filter bank [5, 6]. In the decomposition step in both the DWT and DWPT based coders we have used the Daubechies 9/7 biorthogonal filter bank [7, 8]. The quantizer consists of one uniform scalar quantizer for each subband. For each image the best quantization matrix, containing the quantization steps, is estimated in a rate-distortion sense [4].

We will discuss the lower performance that the modified DWT coder offer in comparison to the modified DCT coder. The resulting insight lead to the development of a modified wavelet packet coder with a new cost function to maximize the influence of the HVS model.

The DWPT coder will result in higher compression in comparison to the two previous works. The cost is more computations both in the encoder and the decoder. Furthermore, only in the DWPT coder the decoding depends on the point of gaze of the observer.

The outline of this paper is as follows. Next, the model of the visual acuity will be described. The proposed scheme is presented in section 3 to 5 followed by simulation results in section 6. Finally, section 7 draws up the final conclusions.
2. A VISUAL ACUITY MODEL

Due to the uneven distribution of cones and ganglion cells in the human retina, we have truly sharp vision only in the central fovea. This covers a visual angle of less than 2 degrees. The ability to distinguishing details is essentially related to the power to resolve two stimuli separated in space. This is measured by the minimum angle of resolution (MAR) [9, 10, 11]. The MAR depends on the eccentricity, which is the angle to the gaze direction. In this work we will use the MAR measured by Thibos [9].

\[ \text{MAR} \]

\[ \text{MSR} \]

\[ \text{focus} \]

\[ \text{origin} \]

\[ \text{observer} \]

Figure 2: Viewing situation.

The related size in the image to a MAR value is called the minimum size of resolution (MSR). This size depends on the current viewing conditions. We will assume that the display is flat. Figure 2 shows the viewing situation.

With a position tracker and an eye tracker we will get the distance between the observer and the display, denoted \( d \), and the point of regard in the image which will be called the focus. From these values, one can calculate the eccentricity, \( e \), for any point, \( p \), in the image plane. Furthermore, the minimum size of resolution for the point \( p \) is equal to,

\[ \text{MSR}_{Lr}(e) = 2 \sqrt{d^2 + r_p^2} \tan \left( \frac{\text{MAR}(e)}{2} \right) \] (1)

where \( r_p \) is the distance between the current point and origin. The MSR in Equation 1 is calculated perpendicular to the \( r_p \) direction. For a computer display the MSR is almost equal for all directions. For larger eccentricities the region which is covered by the MAR will have the form of an oval. However, the \( \text{MSR}_{Lr} \) is used since it is the minimum MSR for all directions. This guarantee that we will not erase visible details.

The MSR bound can be expressed as a visual frequency constraint. Thus, an image frequency must be less than,

\[ f_{\text{VR}}(e) = \frac{1}{2 \cdot \text{MSR}(e)} \] (2)

if an observer is to be able to perceive it.

2.1. Normalized MSR

The model above is a just-noticeable distortion (JND) bound. The possibility to utilize this for compression will occur if we use a large or a high resolution display. However, for many computer displays with normal pixel-resolution the MSR value will be less than the size of a pixel in large parts of the image. If we in these cases want to increase the compression we have to go where \( p_{sw} \) is the width of a pixel. The parameter \( p_{sw} \) is equal to \( p_{sw} = \frac{is}{ir} \) where \( is \) and \( ir \) are the size and pixel resolution of the image. From a JND bound to a MND bound (minimum noticeable distortion). A MND bound can not be measured it has to be estimated. This is achieved by normalization of the JND bound [12].

The normalization will be done such that the relative difficulty of catching details at different eccentricities is kept. This is achieved by scaling the MSR equally at all distances to focus. This is equal to scale the stimulated area on the retina equally at different eccentricities.

We will define the scaling factor so that the normalized MSR is equal to the width of a pixel at the eccentricity \( e_f \) called the fovea-angle. This angle is set to 2 degrees. Thus, the region inside the fovea-angle will be unaffected by the visual model. The normalized MSR is defined as,

\[ \text{MSR}_N(e) = \frac{\text{MSR}(e)}{\text{MSR}(e_f)} \cdot p_{sw} \] (3)

where \( p_{sw} \) is the width of a pixel. The parameter \( p_{sw} \) is equal to \( p_{sw}=is/ir \) where \( is \) and \( ir \) are the size and pixel resolution of the image.

3. EYE-MOVEMENT CONTROLLED TRANSFORM CODER

The DCT, DWT and DWPT schemes will all decompose the image in three signal domains, namely frequency, position and direction. Thus, each transform component represent the energy of the input signal at a certain location in both space and frequency domains.

The energy of a transform component, \( c_i \), is assumed to be located to a corresponding Heisenberg box [5, 6]. A Heisenberg box consist of both an interval in the frequency domain and an interval in the time domain. The lower range of the frequency interval corresponding to component \( c_i \) is denoted \( f_T(c_i) \) and the point in the space interval which is closest to the focus point is denoted \( p(c_i) \).

\[ \text{frequency} \]

\[ \text{space} \]

\[ f_T(c_i) \]

\[ p(c_i) \]

Figure 3: Visual description of the position \( p(c_i) \) and lower frequency range \( f_T(c_i) \) corresponding to component \( c_i \).
The idea in the eye-movement controlled transform coder method is that since we can not perceive high frequencies in the peripheral visual field, we can set corresponding transform components to zero with insignificant loss of visual quality.

According to section 2, for each position at the display we can estimate the visual frequency constraint, \( f_{\text{vc}} \), which is the maximum frequency an observer can perceive at the current position. In addition, each transform component is represented by a frequency, \( f_T \), which is the minimum frequency for which the component will respond and a position which will maximize the \( f_{\text{vc}} \) value. Thus, the strategy above can be expressed as,

\[
f_{\text{vc}}(d, \text{focus}, p(c_i)) < f_T(c_i) \Rightarrow e_i = 0 \quad (4)
\]

where \( d \) is the distance to the observer and \( \text{focus} \) the point on the display which is pointed out by the gaze direction.

4. SPACE-FREQUENCY DECOMPOSITION

To preserve the total number of coefficients the space-frequency resolution is maintained constant. Consider a 1-D signal. The space-frequency decomposition in a DCT and a DWT scheme can be illustrated as in Figure 4. Each coefficient is associated with a Heisenberg box, illustrated as rectangles in the Figures.

![Space frequency decomposition for a DCT scheme (above) and a DWT scheme (below).](image)

Assume that the viewing conditions are equal to, \( d=0.5\text{m}, \sigma=0.3\text{m}, \gamma=512 \text{ pixels}, \epsilon=2 \text{ degrees} \) and focus at the centre of the image. Consider the space frequency decomposition along the positive x-axis. Figure 4 shows this space frequency decomposition and the current visual frequency constraint.

According to Equation 1, a component is set to zero if the corresponding Heisenberg box is totally above the visual constraint in a space-frequency decomposition illustration. Consider figure 4, it is then obvious why the acuity based constraint will set fewer components to zero in the DWT scheme than in the DCT scheme.

To maximize the influence of the visual constraint one would like to have an adaptive space-frequency decomposition which maximize the number of components which are set to zero.

5. MODIFIED DWPT CODER

In a DWPT scheme the decomposition depends on the cost function. Thus, the number of components which will be set to zero by the acuity based constraint will also be dependent on the cost function which has been chosen. If we chose a traditional cost function this number will in some cases be large in some others low.

To maximize the influence of the visual constraint we will define a new cost function. The criterion for the “best” wavelet packet decomposition will be the basis which sets the largest number of components to zero according to the visual constraint. If there exist more than one choice, the decomposition which has the highest space resolution is chosen to minimize the computations and filter spreading. In addition, we will also require that the lowpass branch is decomposed to the maximum depth of the best tree. This since it mostly will increase the compression.

Let a mask, called the VC mask, have the same size as the image and be one at those components which will be kept and zero otherwise. We define the cost function as the sum of the values in the VC mask.

To find the best decomposition,
1. Start with a full uniform wavelet packet decomposition of the maximum allowed depth. Denote this depth with \( D_I \) and set \( i=1 \).
2. Calculate the VC masks for depth \( D_i \) and for depth \( D_{i+1}=D_i-1 \). This is done for each decomposition by calculating a \( f_T \) value for each subband and a \( f_{\text{vc}} \) value for each component and then applying Equation 4.
3. Merge nodes if the cost function decreases or is unchanged. Note that the total cost function is the sum of the cost function for each subband.
4. Iterate \( (i=i+1) \) as long as the cost function is decreased or unchanged.
5. Finally, if necessary split the lowpass branch to the maximum depth of any of the resulting branches.

If no branch in the resulting tree has the maximum allowed depth, the best tree is found. Otherwise, one may
find a better tree if the maximum allowed depth is increased. This is discussed further in section 5.1.

When the best tree is found, the image is transformed according to the final decomposition tree and those transform components which are marked with a zero in the VC mask are set to zero. Note that the structure of the tree does not need to be coded, since the viewing conditions are known by the decoder.

Figure 5 shows a block diagram of the modified DWPT coder.

Figure 5: The modified DWPT encoder.

Since the decomposition tree is dependent on the viewing conditions, a new tree must be calculated each time the focus point is changed.

Figure 6 shows the best wavelet packet decomposition for the same example as in Figure 4. More components are set to zero in the WP decomposition case than for the DCT decomposition case.

Figure 6: Example of a wavelet packet decomposition.

5.1. The Best Tree

This section will discuss the fact that the algorithm above will find the best basis even though not all possible trees are investigated.

The statement is based on two assumptions considering the visual frequency constraint, \(f_{vc}\). First, at the focus point it is larger than the maximum possible frequency in the displayed image. There is no meaning to have a display with higher resolution since the observer will not be able to perceive it. The second assumption is that the visual frequency constraint is a convex function. Both these assumptions are true for all viewing conditions we have examined.

Under these assumptions no component will be set to zero according to the acuity based model for either the decomposition corresponding to maximum frequency or maximum space resolution. Thus, the best tree will instead appear somewhere in between.

Recall that if a component shall be set to zero its corresponding Heisenberg box must be totally above the visual constraint. To decide this we only have to consider the point in each Heisenberg box which is closest to the focus point, denoted \((p, f_T)\). This point will end up in a corner for all boxes except for those which have the focus point inside their space interval.

The algorithm starts with a tree of maximum frequency resolution and will then merge nodes together. Because of the structure of the decomposition scheme only adjacent nodes at the same level in a tree representation can be merged. Furthermore, each node has only one neighbour which it can be merged with. Before a node can be merged with a neighbour, all nodes below the current node must have been merged. Thus, only some of the frequency levels can be removed at a certain decomposition level.

When two nodes in the tree are merged together, all \((p, f_T)\)-points at the frequency level which is removed are exchanged for new points at the nearest lower frequency level. The new points will appear in between the existing points at this level. Figure 7 shows an example.

Figure 7: One decomposition step. Left: tree representation. Right: Heisenberg representation with the \((p, f_{T})\)-points marked as dots.

Let us follow a \((p, f_{T})\)-points through several decomposition steps. For each time it is exchanged the step along the frequency axis will at least be doubled and along the space axis it will at least be divided by two. The result is that the \((p, f_{T})\)-points will be exchanged along concave curves.

Since the visual constraint is a convex function and the \((p, f_{T})\)-points are exchanged along concave curves, the
curve representing a path through the tree can only be above the visual constraint along a connected sequence of points. Thus, there is only one maxima along a curve.

Consider two curves and divide them in before and after there corresponding paths are connected in a common node. After the connecting node, the paths are equal and thereby will the curves be parallel and all existing \((p, f_T)\)-points are at the same frequency level. Before the connecting node, the curves can cross each other. However, below the connection node, the decision to merge nodes and thereby chose \((p, f_T)\)-points along the two paths are separately.

Thus, each path through the tree is represented by a curve and the curve can only be above the visual constraint along a connected sequence of points. As long as the curves can cross each other the chose between the best points are separately. When the chose no longer is separately the curves are parallel. Consequently, the algorithm in the previous section will find the best basis.

6. SIMULATION RESULTS

The modified coder described in this paper will be called the MDWPT coder. The corresponding coder which use the same decomposition but does not set any components to zero according to the acuity based model will be called the DWPT coder.

It is well known that there does not exist any objective distortion measure which completely mirrors the perceived image quality [12]. Furthermore, there are no distortion measurement which consider the acuity of the HVS.

We will instead compare the bit rate for the MDWPT coder and the DWPT coder when they use the same quantization matrix. That way all maintained components will be quantized in the same way and the quality in the fovea region will be equal.

Thus, the procedure has been the following. For a given image we estimate the best quantization matrix in a rate-distortion sense which results in a certain bit rate for the DWPT coder [4]. The image is then coded with this quantization matrix in the two coders. The resulting bit rates are denoted \(R_{DWPT}\) and \(R_{MDWPT}\). We define the compression gain for the modified coder as,

\[
Gain = R_{DWPT}/R_{MDWPT}
\]  

(5)

Figure 7 shows the results when the images Barbara and Lena are coded. The viewing conditions are set to, \(d=0.5m, is=0.3m, ir=512, ef=2\) degrees and focus in the centre.

As can be seen in Figure 6 the gain by the MDWPT coder depends on the required quality. The reason is that if the required quality is low the quantizer in both coders will set the high-pass components to zero anyway. The gain is also dependent on the frequency content in the peripheral parts of the image. The image Lena is smoother than the image Barbara in the peripheral parts and the gain is therefore less for this image.

7. CONCLUSION

The coder described in this paper shows that there is a considerable additional potential for data compression if one takes into account the point of regard of the observer. The gain is dependent on the high frequency content in the peripheral regions of the image and on the quality that is required.

In comparison to the previous works [1, 2] the MDWPT coder will adapt the space frequency decomposition to the current visual frequency constraint and thereby set the highest number of components to zero and achieve the highest compression gain. The cost is that this requires more computations both in the encoder and the decoder.

Our future work will be directed towards investigating the filter spreading outside the Heisenberg boxes. The result could be a better method to predict the effective size of the filters in the space and frequency domain.

Another direction in our future work will be to investigate the real-time performance. Simple visual tests have been done. The observers have then only reported minor artifacts. However, more experiments is necessary too investigate all visual aspects.

A necessary requirement for a system which uses an eye-movement controlled coder is that it can handle the delay introduced by the encoder and the transmission. This is an issue which is not covered in this paper but which will require special attention [13].
REFERENCES


