Visually Lossless Accuracy of Motion Vector in Overcomplete Wavelet-based Scalable Video Coding

Chuan-Ming Song[1], Xiang-Hai Wang[2,3*], Fuyan Zhang[1,3]
[1] Department of Computer Science and Technology, Nanjing University, Nanjing, China
[2] College of Computer and Information Technology, Liaoning Normal University, Dalian, China
[3] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
Email: {chmsong, xhwang}@graphics.nju.edu.cn, fyzhang@nju.edu.cn

Abstract— **In this paper, we propose a visually lossless accuracy model for scalable coding of motion vectors in overcomplete wavelet-based scalable video codec. By exploiting theory of stationary random process, we first estimate motion compensation errors in spatial domain due to inaccurate motion vectors. We then extend the results to overcomplete wavelet domain, and further derive the errors caused by fraction-pixel motion vectors. Finally, combining with a visibility model of wavelet coefficient errors, we propose a novel algorithm to estimate the accuracy threshold of motion vectors with which motion compensation errors will be invisible.**

Experimental results show that our algorithm is effective in estimating visually lossless accuracy threshold of motion vectors. The proposed algorithm can also be used in scalable motion estimation in wavelet domain. It can accelerate motion estimation speed by stopping halfway at the accuracy that will not cause any visible errors.

Index Terms— **scalable video coding, scalable motion vector, motion vector accuracy, human visual system, wavelet**

I. INTRODUCTION

Wavelet-based scalable video coding (WSVC) is an emerging technology that provides video with high adaptability. Many wavelet-based scalable video codecs have been proposed so far [1]–[6]. Nevertheless, when compared with state-of-the-art JSVM [7], the performance of WSVC remains inferior. Thus, further studies are needed to improve its efficiency. In most existing WSVC coders, motion vectors are coded losslessly without bitrate or resolution scalability [8], [9]. This nonscalable manner does not favor the performance at both high bitrates and low bitrates. At low bitrates, motion vectors may occupy an undue portion of available bandwidth. As a result, texture information can not be coded effectively. Moreover, decoded video is always in a reduced quality or resolution. In this case, the degree of tolerance is increased of human visual systems (HVS) to decoded errors. Motion vectors with high accuracy are of little benefit to improve video quality. At high bitrates, motion vectors only consume a minor share of total bandwidth. Texture are decoded with high quality or resolution, in which case HVS becomes more sensitive to poor motion-compensated quality. Motion vectors with high accuracy are needed. Thus, a scalable representation of motion vectors is desirable to balance motion vectors and texture bandwidth over a wide range of bitrates.

Recently, several methods [10]–[20] are proposed for achieving motion vector scalability. These methods can be classified into three categories.

The first category provides multiresolution representation for motion vectors, and exploits the correlation between motion vectors at neighboring scales. Ohm [10] used 3-D Laplacian pyramid to code motion vector field, reducing both spatial and temporal redundancies. Hu [11] and Tsai [12] *et al.* employed quad-tree to partition motion vectors into one based layer and a few enhancement layers in spatial/temporal domain. These vectors are then transmitted from coarse to fine scale. Xiong *et al.* [13] proposed a framework with multi-layer structure of motion vectors. Under this framework they also put forward a decision model to determine appropriate number of layers to be coded into bitstream at a given rate. Secker *et al.* [14] used a linear model to quantify the impact of motion errors on video distortion and designed a scalable coding method using rate-distortion optimized layering technique. Wu *et al.* [15] combined layered structure, alphabet general partition method, and context adaptive binary arithmetic coding to achieve accuracy or quality scalability of motion vectors.

The second category utilizes the spatial dependency be-

This paper is based on "Research on Motion Vector Accuracy in Overcomplete Wavelet-Domain Scalable Video Coding Based on Human Visual System Characteristics" by C.M. Song, X.H. Wang, and F.Y. Zhang, which appeared in the Proceedings of the 9th IEEE Conference on Young Computer Scientists (ICYCS), Zhangjiajie, China, Nov. 2008.

This work was supported by the National Natural Science Foundation of China under Grant Nos. 60372071, 60703084, 60723003, and the Natural Science Foundation of Liaoning Province of China under Grant No.20072156, and the Natural Science Foundation of Jiangsu Province of China under Grant No.BK2007571, and Project Innovation of Graduate Students of Jiangsu Province of China under Grant No.CX07B-121z, and Program for Liaoning Excellent Talents in University under Grant No.RC-04-11.

*Corresponding author.

tween neighboring motion vectors. Barbarien *et al.* [16]–[18] introduced an architecture of quality scalable motion vector coding based on median prediction. Motion vectors are first quantized to form nonscalable base layer. The quantization errors are then coded as enhancement layers using fractional bitplane technique.

The third category estimates motion vectors with different accuracy at different spatial resolution. Boisson [19] and Mrak [20] *et al.* proposed accuracy scalable motion vector coding methods that partitioned bitstream into layers with different accuracy. Motion vectors at lower spatial resolution are decoded with less accuracy, and vice versa.

These methods above effectively enhances performance of spatial, temporal, and rate scalable video coding at low bitrates. Nevertheless, some methods are based on implicit models whose parameters are derived from experiments or empirical results, for example [19]. Such parameters can hardly guarantee that the methods achieve optimal performance. As for other methods,such as [13] and [14], although they set up explicit models in terms of rate-distortion optimization, HVS characteristics are not taken into consideration. If the accuracy of motion vectors are too low on base layer, they can not offer video content a legible representation. As a result, this decoded video will present visible distortions even though the coding method is optimal in terms of objective quality. Thus, scalable coding of motion vectors should be designed on the premise of satisfying HVS requirements.

The major contribution of this paper is a visually lossless accuracy model of motion vectors for overcomplete wavelet-based scalable video codecs. Exploiting theory of stationary random process, we first estimate motion compensation errors in spatial domain due to inaccurate motion vectors. Then we extend the results to overcomplete wavelet domain. Combining with a visual model of wavelet coefficient errors, we propose a novel algorithm for predicting the accuracy threshold of motion vectors with which motion compensation errors will be invisible. This algorithm enjoys the following advantages not shared by conventional methods.

First, the relationship is modeled between motion vector accuracy and motion-compensated distortions in terms of HVS. To the best of our knowledge, few methods have been reported in the literature.

Second, as a generic method, the proposed algorithm can be combined with state-of-the-art scalable coding methods of motion vector. Given resolution and features of video content, our algorithm estimates visually lossless accuracy threshold of motion vectors. As long as the accuracy of base layer is higher than the computed threshold, any current method can then be employed to code motion vectors in the sense of rate-distortion optimization.

Third, our analytical method and results can be extended to other transform domain, such as complete wavelet and contourlet transform, etc., provided that the filter coefficients are given.

The remainder of this paper is organized as follows. Section II models the motion compensation errors due to inaccurate motion vectors. Section III presents estimation method of motion compensation errors in spatial and overcomplete wavelet domain. Section IV describes the proposed visually lossless model of motion vector accuracy and algorithm for computing visually lossless accuracy threshold. Section V reports our experimental results. Finally, Section VI concludes the whole paper.

II. Motion Compensation Errors by Inaccurate Motion Vectors

Currently, block-based motion estimation model is widely used in video coding standards, which assumes that a macroblock in current frame is translation of a macroblock in reference frame. Let $\hat{b}^*_{n+1}(x,y)$ denote ideal prediction of current macroblock B_c, and $\mathbf{V}^* = (v_x, v_y) \in R^2$ be ideal motion vector that minimizes prediction error in terms of mean squared error (MSE). Then the block-based motion estimation can be modeled as

$$\begin{aligned}\hat{b}^*_{n+1}(x,y) &= b_n(x+v^*_x, y+v^*_y) \\ &= b_{n+1}(x,y) + N(x,y) \end{aligned} \quad (1)$$

where $(x,y) \in \{0,1,\ldots,B-1\} \times \{0,1,\ldots,B-1\}$, B is the size of macroblock, and $(n+1)$ is current frame index. $N(x,y)$ denotes interframe noise caused by light changes, camera noise, coding distortions, non-translational motion, occlusions, etc. Without these noises, current macroblock will be identical with its ideal reference macroblock [21]. However, since accuracy and computational complexity have to be balanced, the coded motion vector in practice is a quantized one rather than an ideal one. Supposing that

$$\mathbf{V} - \mathbf{V}^* = (m,n)$$

where $\mathbf{V}$ denotes coded motion vector, the actual prediction of current macroblock is

$$\hat{b}_{n+1}(x,y) = \hat{b}^*_{n+1}(x+m, y+n) + N(x,y). \quad (2)$$

From (2), we can see that prediction of current macroblock is the one resulted from ideal reference macroblock shifted by (m,n). This process is illustrated by Fig. 1, where the macroblock colored in light grey denotes coded reference macroblock, while the one in dark grey denotes the ideal reference. Δ is accuracy of motion vectors.

Based on the above assumption, it is reasonable to believe that motion compensation errors are caused mostly by inaccurate motion vectors. In this case, the prediction error of each pixel in terms of MSE can be calculated by

$$\begin{aligned} e^2 &= E\{|c(x,y) - c(x+m, y+n)|^2\} \\ &= E\{|c(x,y)|^2\} + E\{|c(x+m,y+n)|^2\} \\ &\quad -2E\{|c(x,y)c(x+m,y+n)|\}, \end{aligned} \quad (3)$$

in which $c(x,y)$ denotes pixel intensity at position (x,y), and $E\{\cdot\}$ is expectation operator. Because images can be modeled by a stationary AR-1 process [22] [23], we

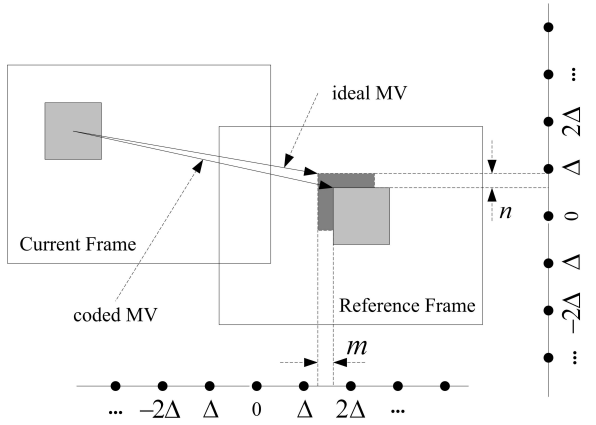

Figure 1. Illustration of the error between coded motion vector and ideal motion vector of a macroblock due to limited accuracy.

assume that pixel intensity in a frame obeys an identical random field. Hence, we approximate (3) by

$$e^2 = 2(\mu^2 + \sigma^2) - 2r(m,n), \tag{4}$$

where μ and σ^2 separately denote the expectation and variance of pixel intensity, and $r(m,n)$ is autocorrelation function of the stationary random process. Definition of $r(m,n)$ is as follows.

$$r(m,n) = E\{c(x,y)c(x+m,y+n)\} \tag{5}$$

Our task now is to compute μ, σ^2, and $r(m,n)$. Then the mean squared motion compensation error (4) will immediately be obtained.

III. Estimation of Motion Compensation Errors

By exploiting theory of stationary random process, we develop an approximation method in this section to compute μ, σ^2, and $r(m,n)$, as well as (4) in spatial and overcomplete wavelet domain.

A. Estimation of Errors in Spatial Domain

For 2-D discrete image and video, Jain [22] provided a method as below used for computing approximately the expectation and variance of pixel intensity.

$$\mu \simeq \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N} c(x,y), \tag{6}$$

$$\sigma^2 \simeq \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}[c(x,y)-\mu]^2, \tag{7}$$

where M and N are height and width of frame, respectively.

Moreover, for a 2-D real random process, its covariance is defined as

$$Cov(m,n) = E\{[s(x,y)-\mu][s(x+m,y+n)-\mu']\} \tag{8}$$

where $s(x,y)$ denotes sample value of (x,y). μ and μ' are expectation of sample values at position (x,y) and $(x+m,y+n)$, respectively. Here we assume that the random process satisfies an identical distribution, then we have

$$r(m,n) = Cov(m,n) + \mu^2 \tag{9}$$

Thus we can compute $r(m,n)$ by $Cov(m,n)$ which is approximated by (10) for digital image and video [22].

$$Cov(m,n) \simeq \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}[c(x,y)-\mu][c(x+m,y+n)-\mu] \tag{10}$$

Using (4)-(10) we can compute the motion compensation errors in pixel domain.

B. Estimation of Errors in Overcomplete Wavelet Domain

Many state-of-the-art wavelet-based scalable video coders carry out motion estimation in wavelet domain. To estimate motion compensation errors for these coders, the method introduced in Section III-B needs to be extended from spatial domain to wavelet domain. One possible method to compute (4) in wavelet domain may be to direct calculate the corresponding μ, σ^2, and $r(m,n)$ of wavelet coefficients. But we have to execute it several times, once for each subband. This is inefficient and expensive. In this section, we model the relationship between motion compensation errors in spatial domain and those in overcomplete wavelet domain. By this model, the motion compensation errors in wavelet domain can be easily computed in terms of μ, σ^2, and $r(m,n)$ in spatial domain. Thus, significant savings of computation can be achieved.

The one-level overcomplete wavelet transform of 2-D images consists of two consecutive steps: filtering horizontally and vertically using a low-pass filter h_0 and a high-pass filter h_1 separately. This process forms four subbands, namely LL_1, LH_1, HL_1 and HH_1, each with the same size as original image. To obtain e^2 in these subbands, we need to compute expectation μ_{sub_1} ($sub_1 \in \{LL_1, LH_1, HL_1, HH_1\}$), variance $(\sigma_{sub_1})^2$, and autocorrelation $r_{sub_1}(m,n)$. Suppose that $c_{sub_1}(x,y)$ is the coefficient of $c(x,y)$ in sub_1 which results from filtering by h_h horizontally and then by h_v vertically. Then we have

$$\begin{aligned}\mu_{sub_1} &= E\{c_{sub_1}(x,y)\} \\ &= E\left\{\left[\sum_{k=1}^{K} h_v(k)\sum_{s=1}^{S} h_h(s)c(x-s,y-k)\right]\right\} \\ &= \mu\sum_{k=1}^{K}\sum_{s=1}^{S} h_v(k)h_h(s),\end{aligned} \tag{11}$$

where K and S are separately the length of filter h_v and h_h. Note that, downsampling operation is not performed in the filtering of overcomplete wavelet transform. Like-

wise, we also have

$$(\sigma_{sub_1})^2 = E\{[c_{sub_1}(x,y) - \mu_{sub_1}]^2\} = \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{t=1}^{T} h_v(k)h_h(s)h_v(j)h_h(t)r(s-t,k-j) - \mu^2 \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{t=1}^{T} h_v(k)h_h(s)h_v(j)h_h(t) \quad (12)$$

and

$$r_{sub_1}(m,n) = E\{c_{sub_1}(x,y)c_{sub_1}(x+m,y+n)\} = \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{t=1}^{T} h_v(k)h_h(s)h_v(j)h_h(t) r(m+s-t,n+k-j), \quad (13)$$

where J and T separately denote the length of filter h_v and h_h.

From (11)-(13) we know that the expectation, variance and autocorrelation of wavelet coefficients can be expressed by convolution of filter bank coefficients with expectation and autocorrelation of pixel intensity.

When wavelet decomposition is further performed on LL_1, the above process can be executed on $c_{LL_1}(x,y)$ to compute μ_{sub_2}, $(\sigma_{sub_2})^2$, and $r_{sub_2}(m,n)$ by substituting $c_{LL_1}(x,y)$ into (11)-(13) and replacing $c(x,y)$. Repeating this process iteratively on $LL_k(k \geq 1)$ at each scale, we then obtain μ_{sub_k}, $(\sigma_{sub_k})^2$, and $r_{sub_k}(m,n)$ for any k-level decomposition. Note that, spatial domain corresponds to the scale space that is one time finer than the scale of sub_1. Set

$$\mu_{sub_0} = \mu \quad \text{and} \quad r_{sub_0}(m,n) = r(m,n).$$

Then the computation of μ_{sub_k}, $(\sigma_{sub_k})^2$, and $r_{sub_k}(m,n)$ can be formulated as

$$\mu_{sub_k} = \mu_{sub_{k-1}} \sum_{k=1}^{K}\sum_{s=1}^{S} h_v(k)h_h(s), \quad (14)$$

$$(\sigma_{sub_k})^2 = \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{t=1}^{T} h_v(k)h_h(s)h_v(j)h_h(t)r_{sub_{k-1}}(s-t,k-j) - (\mu_{sub_{k-1}})^2 \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{t=1}^{T} h_v(k)h_h(s)h_v(j)h_h(t), \quad (15)$$

and

$$r_{sub_k}(m,n) = \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{t=1}^{T} h_v(k)h_h(s)h_v(j)h_h(t) r_{sub_{k-1}}(m+s-t,n+k-j). \quad (16)$$

C. Estimation of Errors by Motion Vectors with Fraction-Pixel Accuracy

Using (4) and (14)-(16), the prediction errors by motion vectors with integer-pixel accuracy are estimated in any subband of overcomplete wavelet. However, for the video of high amount of motion and high detailed texture, motion estimation with fraction-pixel accuracy always significantly improves coding efficiency and decoded quality. Motion compensation errors by fraction-pixel vectors are thus required to be predicted.

Suppose that pixel (coefficient) $c_{sub_k}(x+1/2, y+1/2)(k \geq 0)$ is interpolated by $c_{sub_k}(x,y)$, $c_{sub_k}(x+1,y)$, $c_{sub_k}(x,y+1)$ and $c_{sub_k}(x+1,y+1)$ through bilinear interpolation. Because we assume that all pixels (coefficients) in one frame obey identical distribution, the expectation and variance of $c_{sub_k}(x+1/2,y+1/2)$ are respectively μ_{sub_k} and $(\sigma_{sub_k})^2$. So, only $r_{sub_k}(1/2,1/2)$ is needed to compute e^2. Then we have

$$\begin{aligned} r_{sub_k}(1/2,1/2) &= E\{c_{sub_k}(x,y)c_{sub_k}(x+1/2,y+1/2)\} \\ &= E\{c_{sub_k}(x,y)[c_{sub_k}(x,y) + c_{sub_k}(x+1,y) \\ &\quad + c_{sub_k}(x,y+1) + c_{sub_k}(x+1,y+1)]/4\} \\ &= (\mu_{sub_k}^2 + \sigma_{sub_k}^2)/4 + r_{sub_k}(1,0)/4 \\ &\quad + r_{sub_k}(0,1)/4 + r_{sub_k}(1,1)/4. \end{aligned} \quad (17)$$

Using the same method, we also obtain results as follows.

$$r_{sub_k}(1/2,0) = (\mu_{sub_k}^2 + \sigma_{sub_k}^2)/2 + r_{sub_k}(1,0)/2, \quad (18)$$

$$r_{sub_k}(0,1/2) = (\mu_{sub_k}^2 + \sigma_{sub_k}^2)/2 + r_{sub_k}(0,1)/2, \quad (19)$$

and

$$r_{sub_k}(1/4,1/4) = (\mu_{sub_k}^2 + \sigma_{sub_k}^2)/4 + r_{sub_k}(1/2,0)/4 + r_{sub_k}(0,1/2)/4 + r_{sub_k}(1/2,1/2)/4. \quad (20)$$

From the above results, we further generalize a formula as below which can be used to estimate autocorrelation corresponding to shift of any negative power of two.

$$r_{sub_k}(2^{-n},0) = \sum_{j=1}^{n} 2^{-j}(\mu_{sub_k}^2 + \sigma_{sub_k}^2) + 2^{-n} r_{sub_k}(1,0), \quad (21)$$

$$r_{sub_k}(0,2^{-n}) = \sum_{j=1}^{n} 2^{-j}(\mu_{sub_k}^2 + \sigma_{sub_k}^2) + 2^{-n} r_{sub_k}(0,1), \quad (22)$$

$$r_{sub_k}(2^{-n},2^{-n}) = \sum_{i=1}^{n}\sum_{j=1}^{n} 2^{-(i+j)}(\mu_{sub_k}^2 + \sigma_{sub_k}^2) + \sum_{i=1}^{n} 2^{-(i+n)} r_{sub_k}(1,0) + \sum_{j=1}^{n} 2^{-(j+n)} r_{sub_k}(0,1) + 2^{-2n} r_{sub_k}(1,1). \quad (23)$$

Substituting (23), μ_{sub_k} and $\sigma_{sub_k}^2$ into (4), the motion compensation errors by motion vectors with fraction-pixel accuracy can be estimated in spatial and overcomplete wavelet domain.

Furthermore, from (21)-(23) we can deduce a meaningful conclusion. Let $\Delta_{n,0} = r_{sub_k}(2^{-n},0) - r_{sub_k}(2^{-(n-1)},0)$. Then from (21) we have $\Delta_{n-1,0} - \Delta_{n,0} = 2^{-n}(\mu_{sub_k}^2 + \sigma_{sub_k}^2 - r_{sub_k}(1,0))$. Because $\mu_{sub_k}^2 + \sigma_{sub_k}^2 = r_{sub_k}(0,0) > r_{sub_k}(1,0)$, $\Delta_{n-1,0} - \Delta_{n,0} > 0$ and $\Delta_{1,0} > \Delta_{2,0} > ... > \Delta_{n,0}$. For the other two cases, we can obtain similar results, namely

$\Delta_{0,1} > \Delta_{0,2} > ... > \Delta_{0,n}$ and $\Delta_{1,1} > \Delta_{2,2} > ... > \Delta_{n,n}$. These results indicate that with increase of motion vector accuracy, autocorrelation increases and motion compensation error decreases. Moreover, as the vector accuracy doubles, the increased amplitude $\Delta_{i,0}(i \geq 1)$ (or $\Delta_{0,i}$, $\Delta_{i,i}$) becomes smaller and smaller. It tells us that motion vectors with high accuracy are not the better. With increase of vector accuracy, coding gain gradually declines even to zero. When the accuracy improves from integer pixel to half pixel, the coding gain is most prominent. The conclusion provides a theoretical proof for the experimental results reported in [21].

IV. Visually Lossless Accuracy Model of Motion Vector

The objective of video coding is to approach the ideal rate-distortion curve as much as possible. However, the visual quality should be taken into account as well. Viewers expect scalable video codec to provide a good enough perceptual quality. Hence, we should optimize rate-distortion performance on the premise of satisfying visual quality. In this section, we first propose a visual threshold model of motion vector accuracy according to content characteristics and spatial resolution of video. Then based on this model, we propose a novel algorithm to compute accuracy threshold with which motion vectors will not cause visible motion compensation errors.

A. Wavelet-Based Visually Lossless Threshold Model

Using the model in Section III, we compute motion compensation errors due to inaccurate motion vectors. If the errors exceed a threshold, they will be observed by HVS, reducing subjective quality of decoded video. In this paper, we adopt a visual threshold model proposed by Watson *et al.* [24]. It is used for measuring visual detection threshold of wavelet coefficient error when the coefficient is against complex backgrounds. Its definition is as follows.

$$T_{\lambda,\theta} = \sqrt{D_{\lambda,\theta}^2 + \sigma_{\lambda,\theta}^2}, \tag{24}$$

in which λ denotes decomposition level, and θ is orientation index of subband ($\theta \in \{1,2,3,4\}$, respectively represent LL, HL, HH, LH). $D_{\lambda,\theta}$ is visual threshold for the subband of level λ and orientation θ, expressed in units of the wavelet coefficient. Table I lists the values of $D_{\lambda,\theta}$ for four-level Daubechies 9/7 wavelet when the display visual resolution is 32 pixel/degree. $\sigma_{\lambda,\theta}^2$ denotes variance of corresponding subband.

TABLE I.
Visual Threshold for Four-Level Daubechies 9/7 Wavelet When Display Visual Resolution is 32 Pixel/Degree

	Level			
Orientation	1	2	3	4
1	7.03	5.56	5.68	6.25
2	11.52	7.34	6.36	7.08
3	29.38	14.21	9.77	8.93
4	11.52	7.35	6.36	7.08

As long as motion compensation error e^2 are lower than visual threshold $T_{\lambda,\theta}$, namely

$$e^2 \leq T_{\lambda,\theta}^2, \tag{25}$$

the motion compensated frames will be visually lossless. Thus, our task now is to determine the lowest accuracy of motion vector by which motion compensation errors will not be larger than the computed threshold.

B. Threshold Prediction Algorithm of Visually Lossless Motion Vector Accuracy

Based on the visual threshold model proposed in Section IV-A, we propose a novel algorithm for estimating visually lossless threshold of motion vector accuracy in overcomplete wavelet-based motion estimation.

The motion estimation method adopted in this paper is a multiscale motion estimation [25]. It first constructs an overcomplete representation of reference frame using low-band-shift method. Only the low frequency subbands of reference frames are saved at each scale. Then this method estimates motion vectors of the lowest frequency subband in a band-to-band manner. Afterwards, motion vectors of the low frequency subband at the next finer scale are computed. The prediction of this low frequency subband is decomposed by one-level wavelet to calculate prediction of the high frequency subbands at the coarsest scale. Repeating the above procedure in each low frequency subband from the coarsest scale to the finest one, we will obtain a multiscale prediction of current frame. Since the motion estimation is performed only in low frequency subband at each scale, we substitute $h_h = h_v = h_0$ into (14)-(16) and yield

$$\mu_{LL_k} = \mu_{LL_{k-1}} \sum_{k=1}^{K}\sum_{s=1}^{S} h_0(k)h_0(s) \tag{26}$$

$$(\sigma_{LL_k})^2 = \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{t=1}^{T} h_0(k)h_0(s)h_0(j)h_0(t) r_{LL_{k-1}}(s-t, k-j) - (\mu_{LL_{k-1}}) \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{t=1}^{T} h_0(k)h_0(s)h_0(j)h_0(t) \tag{27}$$

$$r_{LL_k}(m,n) = \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{t=1}^{T} h_0(k)h_0(s)h_0(j)h_0(t) \, r_{LL_{k-1}}(m+s-t, n+k-j) \tag{28}$$

Suppose that components of motion vectors along horizontal direction and vertical direction are with the same accuracy. Then the proposed algorithm for predicting accuracy threshold of motion vectors on the Nth-level of overcomplete wavelet decomposition can be detailed as follows:

1) Select a sub-image B of $P \times Q$ pixels from reference frame, in which P and Q are all integer times of 2^N.

2) Compute μ, σ^2, and r of B using (6)-(10) corresponding to different shift in spatial domain. Let $\mu_{LL_0} = \mu$, $(\sigma_{LL_0})^2 = \sigma^2$, and $r_{LL_0} = r$.
3) Set $\Delta = 1$.
4) Iteratively compute μ_{LL_k}, $(\sigma_{LL_k})^2$, $r_{LL_k}(\Delta, \Delta)$ for $1 \leq k \leq N$ and e^2 of B by (4), (26)-(28). Substitute $(\sigma_{LL_N})^2$ and e^2 into (24)-(25). If (25) satisfies, go to Step 5. Otherwise, go to Step 6.
5) Let $\Delta = 2\Delta$, then compute $r_{LL_N}(\Delta, \Delta)$ and e^2. If (25) still satisfies, then repeat Step 5 until (25) does not satisfy. Then let $\Delta = \Delta/2$, and go to Step 7.
6) Let $\Delta = \Delta/2$, then compute $r_{LL_N}(\Delta, \Delta)$ and e^2 by (4), (23). If (25) does not satisfy, then repeat Step 6 until (25) satisfies. Go to Step 7;
7) Output Δ. The proposed algorithm is over.

Since the maximal error is $\Delta/2$ when a motion vector is quantized by Δ, motion estimation should be implemented with an accuracy not lower than 2Δ in order to ensure visual quality of decoded video.

Note that in the above algorithm, we select a sub-image instead of the whole reference frame for computing visually lossless accuracy threshold. The dimension of the sub-image is smaller than, even equal as well to, that of reference frame. Larger P and Q will produce more accurate threshold, but require higher computational complexity. According to the assumption that pixel intensity in a frame obeys an identical random field, this approach is a reasonable choice to balance computational complexity and prediction accuracy of threshold.

V. Experimental Results and Discussion

To verify the effectiveness of the proposed model, we conduct experiments on standard MPEG-4 test sequences including "Foreman", "Football", "Stefan", and "Coastguard". The wavelet used in our experiments was Daubechies 9/7. Parameters in our algorithm are set as follows: $P = Q = 64$, $N = 2$.

A. Experimental Results

We first compare visual quality of frames motion-compensated by motion vectors with different accuracy in different spatial resolution. Since we find in our experiments that the computed thresholds are all equal or greater than 2, the motion-compensated frames by proposed accuracy are compared only with those by integer accuracy. Fig. 2 and Fig. 3 show comparison results of the 2^{nd} frame of four test sequences in 1/16 CIF resolution and in QCIF resolution, respectively. Meanwhile, the peak signal-to-noise ratio (PSNR) of the above motion-compensated frames is listed in Table II. Although PSNR of the frames motion-compensated by integer vectors is obviously higher than that by vectors with proposed accuracy, these frames present nearly the same subjective quality. Viewing from appropriate distance (23 inches), we can not immediately distinguish the differences between every two corresponding frames in Fig. 2 and Fig. 3 except for "Foreman" sequence.

Since decoded video at low resolution is usually played on devices with low or limited bandwidth, we compress reference frame by Kakadu software [26] at 0.2 bits/pixel to simulate the quality reconstructed at low bitrate. Then we use decompressed frame for motion estimation and compensation. Fig. 4 shows the motion-compensated 2^{nd} frames of "Foreman" and "Football" by proposed accuracy and integer accuracy. Table II also lists the PSNR of these motion-compensated frames. Under such a display condition, from Fig. 4 we can not easily notice the differences between every two corresponding frames, either.

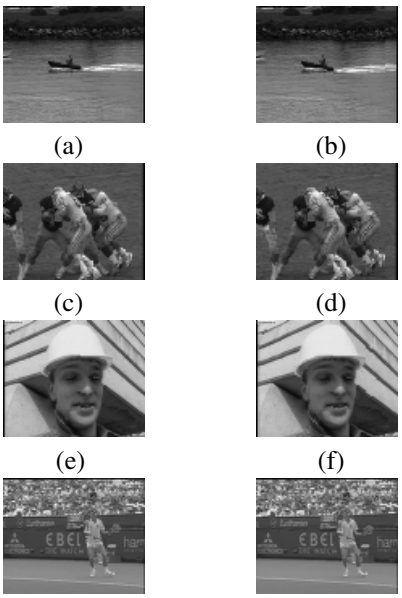

Figure 2. Visual quality comparison of motion-compensated 2^{nd} frames of "Coastguard", "Football", "Foreman" and "Stefan" by motion vectors with integer accuracy (a), (c), (e), (g) and proposed accuracy (b), (d), (f), (h) in 1/16 CIF.

TABLE II.
PSNR Comparison of Motion-Compensated 2^{nd} Frames of Different Sequences by Motion Vectors with Different Accuracy

Resolution	Sequence	Integer accuracy	Proposed accuracy
1/16 CIF	Football	25.14	23.56
	Stefan	27.38	25.85
	Coastguard	29.50	25.23
	Foreman	31.35	29.64
QCIF	Football	25.25	23.63
	Stefan	25.71	22.91
	Coastguard	29.65	26.05
	Foreman	34.15	31.88
QCIF	Foreman (0.2bpp)	32.18	30.35
	Football (0.2bpp)	25.22	23.12

B. Discussion

The proposed model indicates that at low resolution or quality, accurate motion vectors will not lead to improvement of subjective quality. Hence, scalable coding

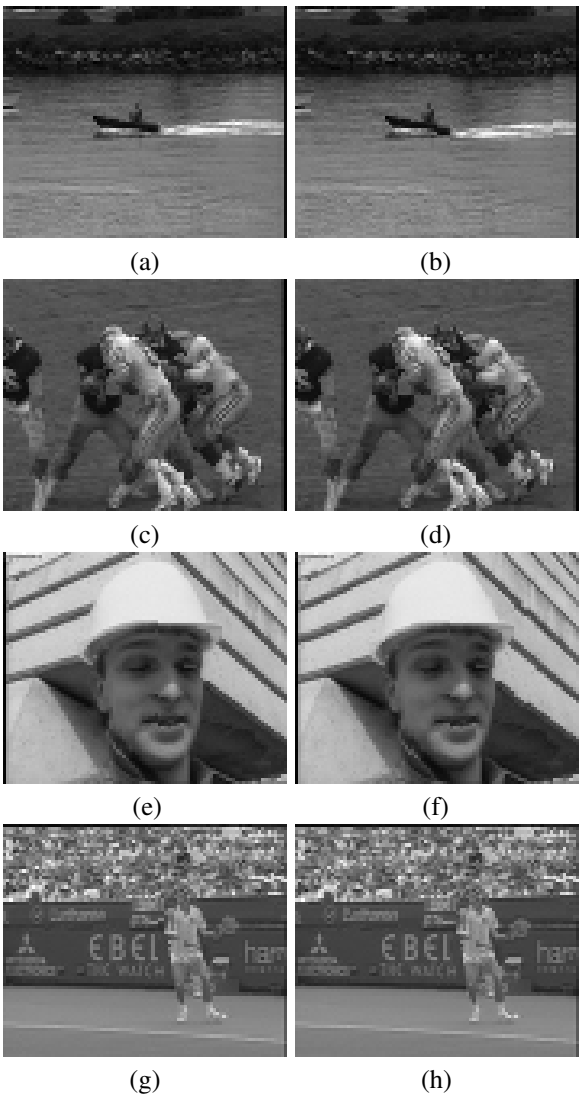

Figure 3. Visual quality comparison of motion-compensated 2^{nd} frames of "Coastguard", "Football", "Foreman" and "Stefan" by motion vectors with integer accuracy (a), (c), (e), (g) and proposed accuracy (b), (d), (f), (h) in QCIF.

of motion vector is not only essential to objective quality improvement of reconstructed frames at low bit rates or resolution, but also straightforward and necessary in the aspect of subjective quality.

Note that the proposed model can be used to estimate accuracy threshold of motion vectors at one scale each time, provided that the vectors at other scales are not quantized. Therefore, if all motion vectors at every scale are quantized by corresponding thresholds at the same time, this model can not estimate exactly motion compensation errors so as to inevitably produce visible mismatch.

In addition, the proposed model does not take temporal masking, frequency masking, and structural correlation into account. This is why Fig. 3(f) presents slightly discontinuous edges. Thus, a more accurate visual model with moderate overhead will be introduced into our model in further study.

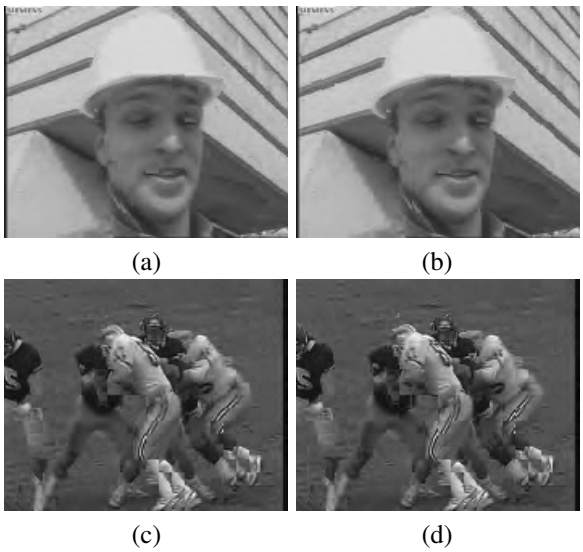

Figure 4. Visual quality comparison of motion-compensated 2^{nd} frames of "Foreman" and "Football" at 0.2bpp by motion vectors with integer accuracy (a), (c) and proposed accuracy (b), (d) in QCIF.

VI. Conclusion

Based on the theory of stationary random process and human visual system model in wavelet domain, this paper constructs a mathematical model for visually lossless accuracy threshold of motion vectors. This model can estimate motion compensation errors due to inaccurate motion vectors with integer and fractional pixel accuracy in spatial domain, as well as in overcomplete wavelet domain. Experimental results support the effectiveness of our model. It can be used for overcomplete wavelet-based scalable motion estimation and scalable coding of motion vectors, enhancing performance of scalable video coders. Moreover, this model can also be extended to estimation of motion compensation errors in other transform domain, provided that the filter coefficients are given.

To the best of our knowledge, few algorithms have been reported on the relationship between motion vector accuracy and visual quality of decoded video in the literature. We believe that our study will be useful in further researches of wavelet-based scalable video coding.

References

[1] Y. Chen and W. A. Pearlman, "Three-Dimensional Subband Coding of Video Using the Zero-Tree Method," in *Proceedings of SPIE Symposiums on Visual Communications and Image Processing*, 1996, pp. 1302–1312.

[2] B. J. Kim, Z. Xiong, and W. A. Pearlman, "Low Bit-Rate Scalable Video Coding with 3D Set Partitioning in Hierarchical Trees (3D SPIHT)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 8, pp. 1374–1387, Aug. 2000.

[3] J. Xu, S. Li, and Y. Q. Zhang, "A Wavelet Coder Using 3-D ESCOT," in *Proc. IEEE-PCM*, 2000.

[4] P. S. Chen and J. W. Woods, "Bidirectional MC-EZBC with Lifting Implementation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 10, pp. 1183–1194, Oct. 2004.

[5] A. Secker and D. S. Taubman, "Lifting-Based Invertible Motion Adaptive Transform (LIMAT)," *IEEE Trans. Image Process.*, vol. 12, no. 12, pp. 1530–1542, Dec. 2003.

[6] Y. Andreopoulos, M. V. der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, "Fully-Scalable Wavelet Video Coding Using In-Band Motion Compensated Temporal Filtering," in *Proceedings of IEEE International Conf. Acoustics, Speech and Signal Processing*, 2003, pp. 417–420.
[7] "Joint Scalable Video Model," JVT of ISO/IEC MPEG & ITU-T VCEG, Tech. Rep. Doc. JVT-X202, 2007.
[8] M. Wien, H. Schwarz, and T. Oelbaum, "Performance Analysis of SVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1194–1203, Sept. 2007.
[9] N. Adami, A. Signoroni, and R. Leonardi, "State-of-the-Art and Trends in Scalable Video Compression with Wavelet-Based Approaches," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1238–1255, Sept. 2007.
[10] J. R. Ohm, "Motion-Compensated 3-D Subband Coding with Multiresolution Representation of Motion Parameters," in *Proceedings of IEEE International Conf. Image Processing*, 1994, pp. 250–254.
[11] Z. Hu, M. van der Schaar, and B. Pesquet-Popescu, "Scalable Motion Vector Coding for MC-EZBC," in *Proceedings of EUSIPCO*, 2004, pp. 657–660.
[12] S. S. Tsai and H. M. Hang, "Motion Information Scalability for MC-EZBC," *Signal Process.: Image Commun.*, vol. 19, no. 7, pp. 675–684, Aug. 2005.
[13] R. Xiong, J. Xu, and F. Wu, "Responses of CE1a in SVC: Scalable Motion," ISO/IEC JTC1/SC29/WG11, Tech. Rep. Doc. M11128, 2004.
[14] A. Secker and D. Taubman, "Highly Scalable Video Compression with Scalable Motion Coding," *IEEE Trans. Image Process.*, vol. 13, no. 8, pp. 1029–1041, Aug. 2004.
[15] Y. Wu and J. W. Woods, "Scalable Motion Vector Coding Based on CABAC for MC-EZBC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 6, pp. 790–795, June 2007.
[16] J. Barbarien, A. Munteanu, F. Verdicchio, Y. Andreopoulos, J. Cornelis, and P. Schelkens, "Scalable Motion Vector Coding," in *Proceedings of IEEE International Conf. Image Processing*, 2004, pp. 1321–1324.
[17] ——, "Prediction-Based Scalable Motion Vector Coding," ISO/IEC JTC1/SC29/WG11, Tech. Rep. Doc. M11016, 2004.
[18] ——, "Motion and Texture Rate-Allocation for Prediction-Based Scalable Motion-Vector Coding," *Signal Process.: Image Commun.*, vol. 20, no. 4, pp. 315–342, Apr. 2005.
[19] G. Boisson, E. Francois, and C. Guillemot, "Accuracy-Scalable Motion Coding for Efficient Scalable Video Compression," in *Proceedings of IEEE International Conf. Image Processing*, 2004, pp. 1309–1312.
[20] M. Mrak, G. C. K. Abhayaratne, and E. Izquierdo, "On the Influence of Motion Vector Precision Limiting in Scalable Video Coding," in *Proceedings of International Conf. Signal Processing*, 2004, pp. 1143–1146.
[21] J. Ribas-Corbera and D. L. Neuhoff, "Optimizing Motion-Vector Accuracy in Block-Based Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 4, pp. 497–510, Apr. 2001.
[22] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[23] J. Liu and P. Moulin, "Information-Theoretic Analysis of Interscale and Intrascale Dependencies Between Image Wavelet Coefficients," *IEEE Trans. Image Process.*, vol. 10, no. 11, pp. 1647–1658, Nov. 2001.
[24] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of Wavelet Quantization Noise," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1164–1175, Aug. 1997.
[25] C. M. Song and X. H. Wang, "A New Scalable Video Motion Estimation Scheme in the Wavelet Domain," *Chinese Journal of Computers*, vol. 29, no. 12, pp. 2112–2118, Dec. 2006.
[26] "Kakadu jpeg2000 software v4.5," available at http://www.kakadusoftware.com/Downloads.html.

Chuan-Ming Song is currently a Ph.D. candidate at Nanjing University, Nanjing, P. R. China. He received his ME and BE degrees both in computer applied technology from the Liaoning Normal University, Dalian, P. R. China, in 2003 and 2006, respectively.

His research interests include scalable image and video coding, and digital watermarking of multimedia.

Xiang-Hai Wang was born in Wangqing, P. R. China. He received his PhD degree in computational mathematics from the Jilin University, Jinlin, P. R. China, in 1999, his MS degree in computer applied technology from the Jilin University in 1995, and his BS degree in mathematics from Jilin Normal University, Siping, P. R. China, in 1986.

He is a Professor at College of Computer and Information Technology, Liaoning Normal University, P. R. China. His current research interests include image and video processing, CG/CAGD, security of multimedia information.

Prof. Wang is a senior member of the China Computer Federation (CCF).

Fuyan Zhang was born in Shaoxing, P. R. China. He received the BS degree in mathematics from Nanjing University, Nanjing, China, in 1962.

He is a Professor at the Department of Computer Science and Technology, Nanjing University. His research interests include multimedia information processing and digital library. On these topics, he has published over 110 papers in international journals and conferences.

Prof. Zhang is a senior member of the China Computer Federation (CCF). He has received many national awards from the Ministry of Education of China.