A FULLY SCALABLE MOTION MODEL FOR SCALABLE VIDEO CODING

Meng-Ping Kao and Truong Nguyen

University of California, San Diego, Department of ECE
http://videoprocessing.ucsd.edu, mkao@ucsd.edu, nguyent@ece.ucsd.edu

ABSTRACT

Motion information scalability is an important requirement for a fully scalable video codec, especially in low bit rate or small resolution decoding scenarios. So far, several layered coding and motion vector precision scalability approaches have been proposed. However, it is still vague on the required functionalities of a fully scalable motion model and how it interacts with other scalabilities, such as spatial, temporal and quality, in a scalable video codec. In this paper, we first define the functionalities required by a fully scalable motion model. Based on these requirements, a fully scalable motion model will be proposed and, moreover, the associated rate distortion optimized estimation techniques will also be provided as a companion. Several simulation results will be presented to summarize the advantages of proposed motion model.

Index Terms— Rate distortion optimization (RDO) motion estimation (ME), scalable motion model, scalable video coding (SVC).

1. INTRODUCTION

Proliferation of high definition displays as well as video enabled devices such as mobile phones has significantly broadened the spectrum of approaches of video coding and processing. Due to dramatic blossom of multimedia applications nowadays, motion pictures will often be transmitted over variable bandwidth channels, both in wireless and cable networks, and be viewed on various display devices, ranging from HDTV to mobile phones. In order to meet a wide range of requirements efficiently, video has to be coded in a more flexible manner such that progressive decoding is feasible for different applications. This is the key incentive of scalable video coding [1], which has been extensively studied in the past few years and is currently undergoing the standardization process held by joint video standard bodies [2].

In general, the three well known scalabilities, i.e. spatial, temporal and quality, are the necessary requirements of a generic SVC, which might also include complexity and error resilience scalabilities. However, these features are not sufficient for a fully scalable codec to be efficient over a broad range of bit-rate. As a core parameter of motion compensated video codec, motion information should also be scalable for flexible adaptations to various decoding scenarios.

Before the definition of scalable motion information is defined in the next section, we would like to review some crucial milestones in this newly emerging field. Quality scalable motion coding is first introduced by Secker [3], in which the wavelet transform coding is applied on motion vector field (MVF) to provide scalability. Maestroni [4], on the other hand, used the quad tree structure bit plane coding technique to encode the variable block size (VBS) MVF. Xiong [5] proposed an estimation algorithm along with his layered scalable motion structure. In his work, an optimal bitstream extractor for decoder is also provided. Mrak [6] has done extensive researches on scalable motion and has proposed both MV accuracy coding and layered motion modeling algorithms.

This paper is organized as follows. Section 2 defines the required functionalities of a fully scalable motion model. We then propose a novel and complete solution for block based motion model in Section 3. The associated RDO ME algorithm is illustrated in more details in Section 4. Section 5 shows the simulation results based on a wavelet SVC codec [7] and the conclusions are given in Section 6.

2. FUNCTIONALITIES OF A FULLY SCALABLE MOTION MODEL

A fully scalable motion model (SMM) is defined to be a single progressively encoded motion bitstream which can be efficiently decoded under any specific spatial, temporal and quality demands from the SVC decoder. The scalable motion bitstream should provide all the information which covers all the decoding possibilities that are available at the decoder. For example, any combination of sequence size ranging from 4CIF to QCIF, frame rate ranging from 30 fps to 7.5 fps, and bit rate ranging from 2000 kbps to 50 kbps should find its corresponding MVF from the scalable motion model, and the most important of all, in a very efficient manner.

From the above definition, it seems the SMM is highly correlated to the SVC codec, i.e. highly codec structure dependent. However, we do find many common properties despite of different codec structures. First, as far as temporal scalability is concerned, the JSVM [2] uses hierarchical B frame structure which is similar to the STAR algorithm [7] we use in our wavelet SVC codec, while the MC-EZBC [8] uses (Unconstrained) MCTF [1] instead. In either case, MVF's associated with the irrelevant frames can be dropped simultaneously to reduce the frame rate with no harm on decoding the remaining frames. As a consequence, temporal scalability of SMM is usually not a problem in most of SVC codec's.

Second, since the true motion should be shared for different resolutions, the spatial scalability of SMM can be easily carried out by down-scaling the highest fidelity MVF according to the desired picture size. For example, the MV's for QCIF sequence are the scaled version of those in CIF sequence. Note that the down-scaling process would cause problems on MV accuracy issue as well as block size issue for a block based motion model. For example, a quarter pixel accuracy MV will result in a one eighth pixel accuracy MV on smaller size pictures which might not be supported by the codec. Moreover, a 4x4 block will become a 2x2 block on smaller size pictures which might also go beyond the decoder capabilities. Therefore, the functionality of spatial scalability in a fully SMM should be formulated as a constraint problem rather than a scalability problem. Our proposed SMM provides full solution on the constraint problem

This work is supported by Conexant Inc. and matching fund from UC Discovery program.

and we will discuss the details in Section 3.

The last property which is also the only explicit scalability SMM should provide is the quality scalability. To be more specific, given a certain motion target bit rate, SMM should be able to provide the best MVF among all possible candidates which occupy no more than the target bit rate. By best we mean under a predefined distortion measurement, i.e. sum of absolute motion compensated difference (SAD). The operating target bit rates could be fine or coarse grain SNR (FGS or CGS) specified, of which FGS is better in terms of more refined optimal operating points.

In general, there are two ways to achieve the quality scalability. The first one is the MV accuracy scalability, which is independent of any underlying motion model. By coding the MV accuracy progressively, i.e. integer MV with half and quarter pixel refinements, the FGS quality scalability can be roughly achieved.

The second way is the motion structure scalability, which is motion model dependent. As a consequence, from now on we will be focusing on block based motion model only. In a block based motion model, VBS tree structure is a common tool to efficiently describe motions in video sequences. By changing the block size, motions of different objects can be better computed. From the VBS point of view, MVF can be decoded in high bit rate with more refined block size or in low bit rate with larger block size and thus fulfills the quality scalability requirement. The combination of both accuracy and motion structure scalability would be ideal for a fully SMM.

Unfortunately, although those scalable functionalities mentioned so far can well depict the structure of an ideal SMM, a corresponding rate distortion optimization process on motion estimation at encoder side and an associated motion bit stream extractor at decoder / transmitter side are needed to ensure overall optimality. Moreover, a more elaborate predictive coding algorithm on motion vector difference (MVD) encoding can also improve the coding efficiency considerably.

3. PROPOSED FULLY SCALABLE MOTION MODEL

Integrating all the above desired features, we propose a novel and fully scalable motion model as shown in Fig. 1. Note that our model has the basic cell as one macro block (MB) so the whole diagram is only the motion structure for one MB. It is clear that we explicitly implement the two refinement dimensions for motion quality scalability in our model, i.e. accuracy and VBS. There's one more concept about VBS scalability we should point out here. As far as scalability is concerned, all the internal nodes in the tree structure should be determined and encoded for possible decoding purposes, as well as the leaf nodes. One example for quad tree structure can be referred to in [4]. To further increase the coding efficiency, an incomplete quad tree structure [9] is adopted in our SMM as shown in Fig. 2. A considerable amount of bits can be saved when some of the leaf nodes have similar MV's to that of their parent node. The decision process is rate distortion optimized as we will discuss in Section 4.

Some notations have to be clarified before further description of our model can proceed. First, we assume that we have A accuracy refinement layers, which are indexed by $a = 0, \cdots, A-1$ with accuracy base layer $a = 0$, and V VBS refinement layers, which are indexed by $v = 0, \cdots, V-1$ with the largest block size layer $v = 0$. We also assume there are R resolution layers, which are indexed by $r = 0, \cdots, R-1$ with the biggest picture $r = 0$.

The motion model shown in Fig. 1 is for resolution $r = 0$, i.e. it contains the highest fidelity motion information for the highest resolution. The descriptions in parenthesis are examples for resolution $r = 0$. For example, "Integer Pel" means that it is the integer pixel accuracy layer for resolution $r = 0$, which should be the half pixel accuracy layer for resolution $r = 1$, and so on. Similarly, "32x32" means that the block size is 32x32 for resolution $r = 0$, which should be 16x16 for resolution $r = 1$, and so on. Therefore, when decoding smaller size pictures, irrelevant information has to be discarded for maintaining high coding efficiency. As mentioned before, the spatial scalability is now posed as two constraints on accuracy and VBS dimensions as follows. Given resolution r, the highest accuracy layer a and the highest VBS layer v are

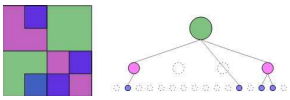


Fig. 2. Incomplete quad tree structure.

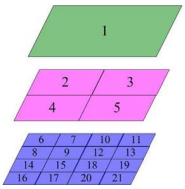


Fig. 3. RDO ME scanning order.

$$a_r = A - 1 - r \quad (1)$$

$$v_r = V - 1 - r \quad (2)$$

respectively. By highest layers a_r and v_r we mean the codec does not support more refined layers for the given resolution r. In other words, quality scalability of SMM can only operate within the range $a = 0, \cdots, a_r$ and $v = 0, \cdots, v_r$, i.e. certain up left corner of the whole SMM diagram.

Knowing the constraints posed by spatial scalability, we are ready to move on to the quality scalability, which is the most important part of SMM. In our SMM, every accuracy refinement layer is associated with a target motion bit rate and is optimized to that bit rate through the RDO process coming out in Section 4. By increasing the total number of accuracy refinement layers, i.e. A, we can be approaching FGS gradually. For example, given a resolution r, the decodable motion quality can be as low as $a = 0$, and can be progressively improved up to $a = a_r$.

The VBS scalability, on the other hand, comes in great effect in a more implicit way as follows. As we refer back to the example in Fig. 1, the incomplete quad tree structure keeps growing as a increases. An intuitive explanation could be that an increasing bit budget for motion model would possibly result in a more refined motion structure as an optimal point on the RD curve. Again, we will discuss the complete RDO process in Section 4.

4. RATE DISTORTION OPTIMIZATION MOTION ESTIMATION

A motion model without the corresponding rate distortion optimization algorithm can not achieve the best coding efficiency. It is the encoder that has full access to the original video sequence and thus should be responsible for providing the best motion information that suits the decoder's requirements. Therefore, given all possible decoding scenarios, a good RDO strategy at the encoder can usually outperforms a good standalone bitstream extractor at the decoder.

In our proposed SMM, the RDO is performed in the basis of sub blocks and the scanning order is shown in Fig. 3 for an example of three VBS layers structure. As observed from Fig. 3, the scanning order is from top layer, $v = 0$, to bottom layer $v = V - 1$, with a raster scan in group of four sub blocks within the same VBS layer.

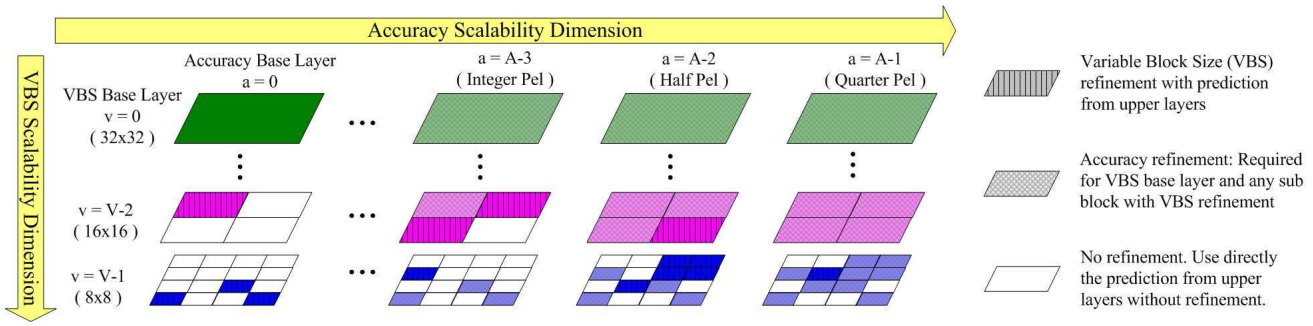


Fig. 1. Proposed fully scalable motion model.

For each sub block, our goal is to determine the best scalable motion vector (SMV) with the following structure in the rate distortion sense.

$$SMV = (a_s, MV_{a_s}, MV_{a_s+1}, \cdots, MV_{A-1}) \tag{3}$$

where a_s is the first accuracy refinement layer for this SMV and MV_a is the refinement information at layer a. Note that $a_s \in \{-1, 0, \cdots, A-1\}$ where $a_s = -1$ denotes no SMV provided. In this case, SMV is degraded as $SMV = (-1)$.

In order to obtain the optimal SMV, the best a_S has to be first determined within the alphabet $\{-1, 0, \cdots, A-1\}$. For this purpose, a new cost function will be introduced. First, a set of rate multiplier is defined as $\{\lambda_a | a = 0, \cdots, A-1\}$ such that each accuracy refinement layer a will be weighted by the penalty multiplier λ_a. In general, smaller a corresponds to lower decoding bit rate, which in turn requires larger penalty multiplier λ_a. Therefore, we usually have the relationship, $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{A-1}$. Given the set $\{\lambda_a\}$, a new rate function is defined as

$$RF(SMV) = \lambda_{a_s} R(a_s) + \sum_{a=a_s}^{A-1} \lambda_a R(MV_a) \tag{4}$$

where $R(\cdot)$ denotes the actual encoding bits function. From the above equation, the rate function of a SMV with up to accuracy layer i can be defined as $RF((a_s, MV_{a_s}, \cdots, MV_i)) = \lambda_{a_s} R(a_s) + \sum_{a=a_s}^{i} \lambda_a R(MV_a)$. Moreover, since the single SMV would provide MV's for all possible resolutions, the total distortion function should also be a combination from all possible resolutions. The individual distortion function from each resolution is weighted by $\{w_r | r = 0, \cdots, R-1\}$. Note that unlike λ_a, w_r has no conventional restrictions on its relative values. Instead, w_r is determined according to the decoder's preference. If the decoder prefers better coding efficiency for a specific resolution r, w_r should be chosen relatively larger than the others, and vice versa. Given the set $\{w_r\}$, a new distortion function is defined as

$$DF(SMV) = \sum_{r=0}^{R-1} w_r D_r(SMV) \tag{5}$$

where $D_r(SMV)$ denotes the distortion measurement when SMV is applied to resolution r. The distortion measurement can be chosen as sum of absolute difference (SAD), sum of square difference (SSD), or reconstruction square error (RSE) [6] for MCTF based SVC. Recall the spatial scalability constraint from (1), $D_r(SMV)$ should be rewritten as $D_r(SMV) = D_r((a_s, MV_{a_s}, \cdots, MV_{A-1-r}))$.

From the above equation, we observe that there is a maximum value for r such that a_r is greater than a_s, i.e. $r_{\max} = A - 1 - a_s$. Define $r_{ub} \triangleq \min(R-1, r_{\max})$ which yields

$$DF(SMV) = \sum_{r=0}^{r_{ub}} w_r D_r((a_s, MV_{a_s}, \cdots, MV_{A-1-r})) \tag{6}$$

Combining (4) and (6), the new cost function can be derived as follows.

$$\begin{aligned} CF(SMV) &= RF(SMV) + DF(SMV) \\ &= \lambda_{a_s} R(a_s) + \sum_{a=a_s}^{A-1} \lambda_a R(MV_a) + \sum_{r=0}^{r_{ub}} w_r D_r((a_s, \cdots, MV_{a_r})) \end{aligned} \tag{7}$$

The RDO process for finding the best SMV is performed layer by layer, i.e. starting from (a_s, MV_{a_s}) and looping through the following refinement layers MV_{a_s+i} incrementally, all the way up to MV_{A-1}. If the best a_s turns out to be -1, all the refinements are automatically set to zeros. Let $SMV_i, i \in \{a_s, a_s+1, \cdots, A-1\}$ denote the current scalable motion vector under estimation whose accuracy refinement layer is up to layer i, i.e. $SMV_i = (a_s, \cdots, MV_i)$ Also let $SMV_i^P = \left(a_s^P, MV_{a_s^P}^P, \cdots, MV_i^P\right), i \in \{a_s, \cdots, A-1\}$ denote the co-located scalable motion vector of SMV_i from the closest upper VBS layer with $a_s^P \leq i$, where a_s^P is the first accuracy refinement layer of SMV_i^P. A slightly modified distortion function for SMV_i^P is now defined as

$$DF\left(SMV_i^P\right) = \sum_{r=0}^{r_{ub}} w_r D_r\left((a_s^P, MV_{a_s^P}^P, \cdots, MV_i^P)\right) \tag{8}$$

which measures the weighted distortion throughout the same resolutions as SMV_i does. The pseudo code for finding the best pair is now listed as follows:

For $a_s = 0 : A-1$
  $MV_{a_s} = \arg\min_{MV}(CF(a_s, MV))$
  If $CF(a_s, MV_{a_s}) = CF(SMV_{a_s}) < DF\left(SMV_{a_s}^P\right)$
    Break
  Elseif $a_s == A-1$
    $(a_s, MV_{a_s}) = (-1, 0)$

Once $a_s \neq -1$, $MV_i, i = a_s+1, \cdots, A-1$ are sequentially determined according to the following rule.

$$MV_i = \arg\min_{MV}(CF(a_s, MV_{a_s}, \cdots, MV_{i-1}, MV)) \tag{9}$$

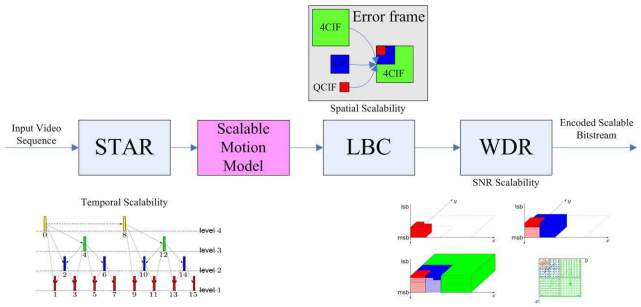


Fig. 4. SVC system diagram with proposed SMM imbedded.

Table 1. Maximum Bit Rate Allocations for SVC Encoder (kbps)

	30 fps	15 fps	7.5 fps
CIF	1536	768	-
QCIF	512	256	128

5. SIMULATION RESULTS

The evaluation of our proposed SMM will be performed on the low complexity wavelet based SVC framework [7]. Fig. 4 shows the system diagram with the insertion of scalable motion model. Note that this SVC framework can be classified as 2D+t+2D [1], which best suits our SMM and the associated scalable motion estimation algorithm.

The format of input testing sequence is CIF with 30 fps and the SVC will generate the scalable bitstream with maximum bit rates for various decoding scenarios as listed in Table 1. Our SMM will be compared side by side with non-scalable motion model in both CIF and QCIF formats.

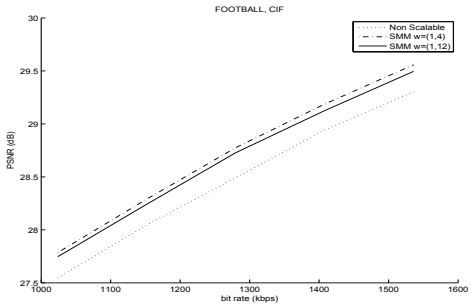
The rate distortion curve for FOOTBALL sequence in CIF size is shown in Fig. 5(a). Here we try two different settings on w_r, where $w = (1, 4)$ puts equal weightings on both CIF and QCIF sizes while $w = (1, 12)$ puts more weightings on QCIF size sequence. It is clear that both settings outperform the non-scalable motion model. Moreover, $w = (1, 4)$ yields better coding efficiency on CIF sequence as expected. This result verifies that our SMM has the ability to fine tune the coding performance toward a preferred resolution using the distortion multiplier w_r. The corresponding tradeoff would be degradations on other resolutions, as shown in Fig. 5(b) for QCIF size sequence.

6. CONCLUSIONS

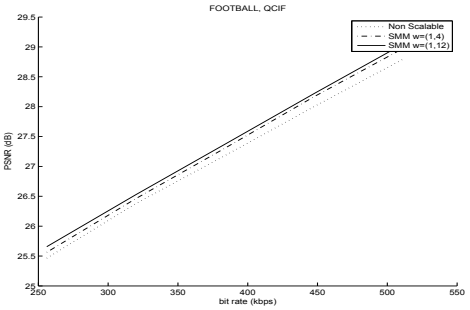
A novel and fully scalable motion model has been proposed to enable a scalable video codec achieving optimality over a wide range of bit rate, resolution and frame rate. The associated rate distortion optimization algorithm provides the tool, via the new introduced rate and distortion multipliers, to further optimize the coding efficiency toward a preferred decoding scenario, with minimal degradation on others. Simulations have shown promising results that verify the various functionalities of our SMM.

7. REFERENCES

[1] Jens-Rainer Ohm, "Advances in scalable video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 42–56, Jan. 2005.



(a)



(b)

Fig. 5. Comparison of RD curves between non-scalable motion model (dotted line) and proposed SMM (dashed line with $w = (1, 4)$ and solid line with $w = (1, 12)$) using FOOTBALL as input sequence. (a) CIF size sequence. (b) QCIF size sequence.

[2] ISO/IEC JTC1/SC29/WG11, "Joint scalable video model JSVM-9," 79th MPEG Meeting, Marrakech, Morocco, JVT-U202, Jan. 2007.

[3] Andrew Secker and David Taubman, "Highly scalable video compression with scalable motion coding," *IEEE Trans. Image Process.*, vol. 13, no. 8, pp. 1029–1041, Aug. 2004.

[4] Davide Maestroni, Aitgirsto Sarti, Marco Tagliasacchi, and Stefano Tubaro, "Scalable coding of variable size blocks motion vectors," in *Proc. Int. Conf. Image Process.*, 2004, vol. 2, pp. 1333–1336.

[5] Ruiqin Xiong, Jizheng Xu, Feng Wu, Shipeng Li, and Ya-Qin Zhang, "Layered motion estimation and coding for fully scalable 3D wavelet video coding," in *Proc. Int. Conf. Image Process.*, 2004, vol. 4, pp. 2271–2274.

[6] Marta Mrak, Nikola Sprljan, and Ebroul Izquierdo, "Evaluation of techniques for modeling of layered motion structure," in *Proc. Int. Conf. Image Process.*, 2006, pp. 1905–1908.

[7] Meng-Ping Kao and Truong Nguyen, "Motion vector field manipulation for complexity reduction in scalable video coding," in *40th Asilomar Conf. Signals Syst. Comput.*, 2006, pp. 1095–1098.

[8] Shih-Ta Hsiang and John W. Woods, "Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank," *Signal Process.: Image Commun.*, vol. 16, pp. 705–724, 2001.

[9] Jong Won Kim and Sang Uk Lee, "On the hierarchical variable block size motion estimation technique for motion sequence coding," *SPIE Visual Commun. Image Process.*, 1993.