An Optimal Quad-Tree-Based Motion Estimator

[*]Guido M. Schuster and [†]Aggelos K. Katsaggelos

Northwestern University, Department of Electrical and Computer Engineering
Evanston, IL 60208, Email: [*]gmschust@ece.nwu.edu, [†]aggk@ece.nwu.edu

ABSTRACT

In this paper we propose an optimal quad-tree (QT)-based motion estimator for video compression. It is optimal in the sense that for a given bit budget for encoding the displacement vector field (DVF) and the QT segmentation, the scheme finds a DVF and a QT segmentation which minimizes the energy of the resulting displaced frame difference (DFD). We find the optimal QT decomposition and the optimal DVF jointly using the Lagrangian multiplier method and a multilevel dynamic program. The resulting DVF is spatially inhomogeneous since large blocks are used in areas with simple motion and small blocks in areas with complex motion. We present results with the proposed QT-based motion estimator which show that for the same DFD energy the proposed estimator uses about 30% fewer bits than the commonly used block matching algorithm.

KEYWORDS: Motion Estimation, Video Compression, Operational Rate Distortion Theory, Quad-Tree, Lagrangian Relaxation, Dynamic Programming

1 INTRODUCTION

Since the bandwidth of uncompressed video is usually much larger than the available channel capacity, video compression is one of the enabling technologies behind the multimedia revolution. Video lends itself to compression because of the high perceptual, temporal and spatial redundancies inherent in a natural scene. The most common approach to exploit the temporal redundancy is motion compensated prediction. All current video coding standards such as MPEG-1, MPEG-2, H.261 and H.263 are motion compensated video coders, where the current frame is predicted using a previously reconstructed frame and motion information, which needs to be estimated. The motion estimation problem for video coding differs from the general motion estimation problem in the following ways. First, for video coding, the performance of a motion estimator can be assessed using a distortion measure, whereas for the general motion estimation problem, such measures are very hard to define. Second, for video coding, the DVF has to be transmitted and therefore the minimization of the distortion should be constrained by the available number of bits to encode the DVF, whereas for the general motion estimation problem, the constraints are used to enforce certain desirable properties of the DVF, such as smoothness. Third, for video coding, the resulting DVF can be arbitrary, i.e., it does not have to correspond to the real motion in the scene, as long as the distortion is minimized for the available bit rate.

This paper is organized as follows: In Sec. 2 we formulate the problem and discuss some of the previous solution attempts. In Sec. 3 we introduce the necessary notation and the underlying assumptions. In Sec. 4 we formulate the rate constrained motion estimation problem and show how the Lagrangian multiplier method can be used to find the optimal solutions on the convex hull of the operational rate distortion curve. In Sec. 5 we develop a fast algorithm to solve the relaxed problem using dynamic programming (DP). In Sec. 6 we compare the proposed QT-based motion estimator with block matching and in Sec. 7, we summarize the paper.

2 PROBLEM FORMULATION

All current video coding standards are based on a fixed segmentation of the frame into blocks of a given size. The main advantages of this are simplicity and robustness of the algorithm and the fact that no segmentation information needs to be transmitted. The only exception to this is H.263 when the "Advanced Prediction Mode" is used. Then there is the option to split the 16×16 block into four 8×8 blocks. The main disadvantage of a fixed and arbitrary segmentation is that it can not accurately represent real objects, and hence the representation of the scene is not as compact as it could be. In addition, the underlying motion model of simple translation cannot capture more complex motion, such as, pan, rotation and zooming. Clearly a dense DVF can represent more complicated motion. The problem, however, in this case is that in many regions of the scene, most of the motion vectors are identical and therefore not needed to describe the motion accurately. Hence a natural DVF is inherently inhomogeneous, in the sense that for regions with simple motion, one motion vector is sufficient, whereas in regions with complex motion the DVF should be denser. If the scene is segmented into blocks of different sizes, then a compact representation of the DVF can be achieved. Since the segmentation information of the frame needs to be transmitted, an efficient method should be used to encode it.

The QT data structure is commonly used to decompose a given frame into blocks of different sizes, since it enables an efficient representation of the resulting decomposition. A QT starts with a square block with side length a power of two. This block can be split into four equally sized square sub-blocks, and only one bit is required to encode the split or lack of it. Clearly this splitting can then be recursively applied to the sub-blocks until a sub-block is of dimensions 1×1. Hence the entire segmentation of the frame can be represented by a tree structure, where every node has four children. The QT structure is an efficient way of segmenting the frame into blocks of different sizes. It is therefore very attractive for the efficient representation of an inhomogeneous DVF. In Refs.,[1–3] the QT structure is used to compactly represent the DVF. In Ref.[1] the dense motion field is represented efficiently using higher order motion models, and the spatial density of the applicability of these models is encoded using a QT. In Ref.[2] the temporal gradient of a pel-recursive motion estimation algorithm is QT encoded and in Ref.[3] a QT-based motion estimator is proposed which finds the best QT in the rate distortion sense (note that this algorithm is a direct application of the more general scheme proposed in Ref..[4]) The motion estimator we propose is more general and more efficient than the scheme in Ref.,[3] since it finds the optimal QT decomposition and the optimal DVF, given that the DVF is first order differential pulse code modulation (DPCM) encoded. The presented algorithm is a generalization of the scheme presented in Ref.,[4] since the optimal QT decomposition can be found even when there are dependencies between the QT leafs.

Note that the variable block size based decomposition of the frame is a good compromise between the overly complex object-oriented approaches,[5] and the overly simple fixed block size based schemes. The object-oriented approaches suffer from the fact that it is in general very hard to accomplish joint segmentation and motion estimation, while the fixed block size schemes suffer from the inability to represent complex motion. In addition, the overhead paid for the segmentation of a QT-based scheme is much lower than the bit rate required to transmit the fine segmentation of an object-oriented scheme,[6] although it is still substantial considering the fact that a fixed segmentation based scheme does not require any segmentation information to be transmitted.

3 NOTATION AND ASSUMPTIONS

Consider Fig. 1, where a frame is segmented using variable block sizes and the black curve indicates the feature of interest. The QT shown in Fig. 2 is used to represent this segmentation. As can be seen in this figure, the QT data structure decomposes a $2^N \times 2^N$ image (or block of an image) down to blocks of size $2^{n_0} \times 2^{n_0}$. This decomposition results in an $(N - n_0 + 1)$-level hierarchy $(0 \leq n_0 \leq N)$, where all blocks at level n $(n_0 \leq n \leq N)$ have size $2^n \times 2^n$. This structure corresponds to an inverted tree, where each $2^n \times 2^n$ block (called a *tree node*) can either be a *leaf*, i.e., it is not further subdivided, or can branch into four $2^{n-1} \times 2^{n-1}$ blocks, each a *child* node. The tree can be represented by a series of bits that indicate termination by a leaf with a "0" and branching into child nodes with an "1". Let $b_{l,i}$ be the block i at level l, and the children of this block are therefore $b_{l-1,4*i+j}$, where $j \in [0,1,2,3]$. The complete tree is denoted by $\mathcal{T}$ and a tree node is identified by the ordered pair (l, i) (see Fig. 3). This ordered pair is called the index of the tree node. Each leaf of $\mathcal{T}$ represents a particular block in the segmented frame. For future convenience, let the leafs be numbered from zero to the total number of leafs in the QT $(N_{\mathcal{T}})$ minus one, from left-to-right and hence in increasing order of the measure $4^{l-n_0} * i$ (this ordering of the leafs is indicated by italic numbers in Fig. 3). Let $m_{l,i} \in M_{l,i}$ be the motion vector for block $b_{l,i}$, where $M_{l,i}$ is the set of all admissible motion vectors for block $b_{l,i}$. Let $s_{l,i} = [l, i, m_{l,i}] \in S_{i,l} = \{l\} \times \{i\} \times M_{l,i}$ be the local state for block $b_{l,i}$, where $S_{l,i}$ is the set of all admissible state values for block $b_{l,i}$. Let $x = [l, i, m] \in X = \bigcup_{l=N}^{n_0} \bigcup_{i=0}^{4^{N-l}-1} S_{l,i}$ be the global state and X the set of all admissible state values. Finally let $x_0, \ldots, x_{N_{\mathcal{T}}-1}$ be a global state sequence, which represents the left-to-right ordered leaves of a valid QT $\mathcal{T}$.

The DFD energy $D(x_0, \ldots, x_{N_{\mathcal{T}}-1})$ is the sum of the individual block energies $d(x_j)$, that is,

$$D(x_0, \ldots, x_{N_{\mathcal{T}}-1}) = \sum_{j=0}^{N_{\mathcal{T}}-1} d(x_j). \qquad (1)$$

This is true since each block in the current frame is predicted using one motion vector which points to a block in a previously reconstructed frame. Note that in the rest of the paper we will use the terms *DFD energy* and *distortion* between the original and predicted frames, interchangeably.

As pointed out in the introduction, the DVF exhibits large spatial correlation, and hence we encode it using a first order DPCM. In other words, we allow for a first order dependency between the QT-leaves along the scanning path. Because of the variable block sizes, we add the constraint that blocks which belong to the same parent, need to be scanned in sequence. We discuss the selection of a good scanning path in more detail in Sec. 3.1. Based on the first order DPCM assumption, the DVF rate $R(x_0, \ldots, x_{N_{\mathcal{T}}-1})$ can be expressed as follows,

$$R(x_0, \ldots, x_{N_{\mathcal{T}}-1}) = \sum_{j=0}^{N_{\mathcal{T}}-1} r(x_{j-1}, x_j), \qquad (2)$$

where $r(x_{j-1}, x_j)$ is the block rate which depends on the motion vectors of the current and previous blocks, since the difference between these two motion vectors is entropy encoded.

3.1 Scanning path based on Hilbert curves

A scanning path is a rule which defines in what order the different blocks of a frame are visited. Since the QT decomposition can change from frame to frame, each frame might require a different scanning path. We propose a procedure for inferring the scanning path from a given QT decomposition. This procedure, together with an appropriate level n_0 scanning path, guarantees that consecutive blocks are always neighboring blocks, that is, they have a common edge. This results in an efficient encoding of the DVF using a first order DPCM scheme along the scanning path.

Assume that the frame is completely segmented into blocks of the smallest size, i.e., $2^{n_0} \times 2^{n_0}$. The QT representation of this segmentation is a complete tree, where all leafs are of level n_0. Further assume that the scanning path for this decomposition is known to the encoder and decoder. The scanning path for any other QT decomposition is defined recursively, by merging four consecutive blocks along the scanning path at the lower level to form the blocks at the higher level. An example of this recursive definition is given in Fig. 8, where the level n_0 scanning path is a third order Hilbert curve.[7] From the properties of a Hilbert curve and the above procedure it is clear that when the level n_0 scanning path is a Hilbert curve of order $N - n_0$, then the overall scanning path connects only neighboring blocks and the resulting blocks are square, regardless of the QT decomposition. Since for our experiments we selected a quarter common intermediate format (QCIF) sequence which is of dimension 176×144, the level n_0 scanning path cannot be a pure Hilbert curve. Figure 5 shows the modified Hilbert scan used in the experiments, where the smallest block size is 8×8 ($n_0 = 3$). Note in Fig. 5 that the lower left corner ($x = [0, \ldots, 127], y = [16, \ldots, 143]$) of the level n_0 scanning path is a pure Hilbert curve of order 7 ($2^7 = 128$).

4 RATE DISTORTION FORMULATION

The motion estimation problem for video coding can be stated as the following constrained optimization problem,

$$\min_{x_0,\ldots,x_{N_{\mathcal{T}}-1}} D(x_0, \ldots, x_{N_{\mathcal{T}}-1}), \qquad \text{subject to:} \qquad R(x_0, \ldots, x_{N_{\mathcal{T}}-1}) \leq R_{max}, \tag{3}$$

where $D(x_0, \ldots, x_{N_{\mathcal{T}}-1})$ and $R(x_0, \ldots, x_{N_{\mathcal{T}}-1})$ are defined by Eqs. (1) and (2), respectively. In other words, we try to find the inhomogeneous DVF which results in the smallest DFD energy for a given maximum bit rate for encoding the DVF. Clearly the optimal state sequence $x_0^*, \ldots, x_{N_{\mathcal{T}}-1}^*$ identifies the optimal QT decomposition and the optimal DVF, since each state value x_j contains the block size, its location and the associated motion vector.

We propose to use the Lagrangian multiplier method[8] to solve the constrained problem of Eq. (3). The idea behind this method is to transform the "hard" constrained optimization problem into a family of "easy" unconstrained problems, which can be solved efficiently. This transformation is achieved by creating a new objective function, the Lagrangian cost function ($J_\lambda(\cdot)$). It is the sum of the original objective function and the constraint, where the constraint is weighted by the Lagrangian multiplier λ,

$$J_\lambda(x_0, \ldots, x_{N_{\mathcal{T}}-1}) = D(x_0, \ldots, x_{N_{\mathcal{T}}-1}) + \lambda * R(x_0, \ldots, x_{N_{\mathcal{T}}-1}). \tag{4}$$

The main theorem[8] of the Lagrangian multiplier method states that if there is a λ^* such that

$$[x_0^*, \ldots, x_{N_{\mathcal{T}}-1}^*] = \arg \min_{x_0,\ldots,x_{N_{\mathcal{T}}-1}} J_{\lambda^*}(x_0, \ldots, x_{N_{\mathcal{T}}-1}) \tag{5}$$

leads to $R(x_0, \ldots, x_{N_{\mathcal{T}}-1}) = R_{max}$, then $x_0^*, \ldots, x_{N_{\mathcal{T}}-1}^*$ is also an optimal solution to (3). The main shortcoming of the Lagrangian multiplier method comes from the fact that only solutions which belong to the convex hull of the operational rate distortion curve (or in this case, the rate energy curve) can be found. Since for the proposed scheme, the convex hull solutions tend to be dense (see Sec. 6), this is not a problem in practice. Note that the dual problem, which can be stated as follows,

$$\min_{x_0,\ldots,x_{N_{\mathcal{T}}-1}} R(x_0, \ldots, x_{N_{\mathcal{T}}-1}), \qquad \text{subject to:} \qquad D(x_0, \ldots, x_{N_{\mathcal{T}}-1}) \leq D_{max}, \tag{6}$$

can be solved with exactly the same technique using the following relabeling of function names,
$R(x_0, \ldots, x_{N_{\mathcal{T}}-1}) \leftarrow D(x_0, \ldots, x_{N_{\mathcal{T}}-1})$, $D(x_0, \ldots, x_{N_{\mathcal{T}}-1}) \leftarrow R(x_0, \ldots, x_{N_{\mathcal{T}}-1})$.

Having stated the main theorem of the Lagrangian multiplier method, there are two problems left to address: first, how to find the optimal λ^* of Eq. (5) and second, how to solve the unconstrained problem of Eq. (5) optimally for an arbitrary λ. In Ref.[9] we proposed a very fast convex search to find λ^* and in Sec. 5 we introduce a dynamic programming algorithm to find the optimal solution to problem (5).

5 OPTIMAL SOLUTION OF THE UNCONSTRAINED PROBLEM BY MULTILEVEL DYNAMIC PROGRAMMING

The form of the objective function of the optimization problem of Eq. (5) suggests that dynamic programming (DP) should be used to find the optimal solution efficiently. To be able to employ forward DP (the Viterbi algorithm), a DP recursion formula needs to be established. A graphical equivalent of the DP recursion formula is a trellis where the admissible nodes and the permissible transitions are explicitly indicated. Consider Fig. 6 which represents the multilevel trellis for a 32×32 image block ($N = 5$), with a QT segmentation developed down to level 3 ($n_0 = 3$, block size 8×8). The QT structure is indicated by the white boxes with the rounded corners. These white boxes are not part of the trellis used for the Viterbi algorithm but indicate the set of admissible state values $S_{l,i}$ for the individual blocks $b_{l,i}$. The black circles inside the white boxes are the nodes of the trellis (i.e., the state values $s_{l,i}$). Note that for simplicity, only two trellis nodes per QT node are indicated, but in general, a QT node can contain any number of trellis nodes. The auxiliary nodes, start and termination (S and T) are used to initialize the DPCM and to select the path with the smallest cost.

Each trellis node represents the prediction of the block it is associated with using a different motion vector. Since the state of a block is defined to contain its level and number within that level (which identifies the blocks size and its position in the frame), and its motion vector, each of the nodes contains the distortion occurring when the associated block is predicted using the node's motion vector. As can be seen in Fig. 6, not every trellis node can be reached from every other trellis node. By restricting the permissible transitions, we are able to force the optimal path to select only valid QT decompositions. Such valid decompositions are based on the fact that at level l, block $b_{l,i}$ can replace four blocks at level $l-1$, namely $b_{l-1,4*i+0}, b_{l-1,4*i+1}, b_{l-1,4*i+2}$ and $b_{l-1,4*i+3}$. As we will see later in this section, the QT encoding cost can be distributed recursively over the QT so that each path picks up the right amount of QT segmentation overhead.

Assume that no QT segmentation is used and the block size is fixed at 8×8. In this case, only the lowest level of the trellis in Fig. 6 is used. The transition costs between the trellis nodes would be the rate required to encode the differences between consecutive motion vectors along the scanning path weighted by the Lagrangian multiplier λ. Assume now that the next higher level, level 4, of the QT is included. Clearly the transition cost between the trellis nodes of level 3 stay the same. In addition, there are now transition costs between the trellis nodes of level 4 and also transition costs from trellis nodes of level 3 to trellis nodes of level 4 and vice versa, since each cluster of four blocks at level 3 can be replaced by a single block at level 4. The fact that a path can only leave and enter a certain QT level at particular nodes results in paths which all correspond to valid QT decompositions. Note that every QT node in a path is a leaf of the QT which is associated with this path. In this example, a tree of depth 3 has been used to illustrate how the multilevel trellis is built. For QTs of greater depth, a recursive rule has been derived which leads to the proper connections between the QT levels.[10] In the presented multilevel trellis, the nodes of the respective blocks hold the information about the distortion occurring when the associated block is predicted using the motion vector of the node. The rate needed to encode the difference between the motion vectors is incorporated into the transition cost between the nodes (weighted by the Lagrangian multiplier λ), but so far, the rate needed to encode the QT decomposition has not been addressed.

Since the Viterbi algorithm will be used to find the optimal QT decomposition, each node needs to contain a term which reflects the number of bits needed to split the QT at its level. Clearly, trellis nodes which belong to blocks of smaller size have a higher QT segmentation cost than nodes which belong to bigger blocks. When the path includes only the top QT level N, then the QT is not split at all, and only one bit is needed to encode this. Therefore its segmentation cost, $A_{N,0}$, equals one. For the general case, if a path splits a given block $b_{l,i}$ then a segmentation cost of $A_{l,i} + 4$ bits has to be added to its overall cost function, since by splitting block $b_{l,i}$, 4 bits will be needed to encode whether the four child nodes of block $b_{l,i}$ are split or not. Since the path only visits trellis nodes and not QT nodes, this cost has to be distributed to the trellis nodes of the child nodes of block $b_{l,i}$. How the cost is split among the child nodes is arbitrary since every path which visits a sub-tree rooted by one child node, also has to visit the other three sub-trees rooted by the other child nodes. Therefore the path will pick up the segmentation cost, no matter how it has been distributed among the child nodes. Since the splitting of a node at level $n_0 + 1$ leads to four child nodes at level n_0, which can not be split further, no segmentation cost needs to be distributed among its child nodes. Clearly, using the above argument, these segmentation costs can be generated recursively in a Top-Down fashion.

The recursion involved in the assignment of the encoding cost is illustrated in Fig. 7. Note that in Fig. 7, the segmentation cost is distributed along the leftmost child. As mentioned before, any other assignment of the segmentation cost will lead to the same result. Furthermore, since in the Lagrangian cost function the rate and distortion are merged by adding the rate, weighted by the Lagrangian multiplier, to the distortion, the segmentation rate also needs to be weighted by λ.

Having established the multilevel trellis, the forward DP algorithm can be used to find the optimal state sequence $x_0^*, \ldots, x_{N_\mathcal{T}-1}^*$ which will minimize the unconstrained problem (5). The Viterbi algorithm simply finds the shortest path from S to T, where the distance is measured as the sum of the node distortions $d(x_j)$ plus the sum of the weighted segmentation and DVF encoding rates, $\lambda * r(x_{j-1}, x_j)$. Hence the Viterbi algorithm finds the optimal solution to the unconstrained problem (5). In Fig. 7, a QT of depth 4 is displayed and the optimal state sequence is indicated which leads to the segmentation shown in Fig. 8. Note that the resulting scanning path is spatially non-disruptive and the segmentation cost along the optimal path adds up to 13 bits, which is the number of bits needed to encode this QT decomposition. The bit stream for this QT decomposition is “1010000011001”.

5.1 Color

In the our experiments, we only use the luminance part of the sequence, but the presented theory also covers the case when the chrominance distortion is included in the distortion measure. The block distortion $d(x_j)$ can be defined arbitrarily and hence it can contain contributions from the luminance channel as well as from the two chrominance channels (or from R, G and B channels). In the most general form, the block distortion can be written as follows,

$$d(x_j) = \phi(d^Y(x_j), d^{Cb}(x_j), d^{Cr}(x_j)), \tag{7}$$

where $\phi(\cdot)$ is an arbitrary function and the superscripts Y, Cb and Cr indicate the distortions in the luminance and chrominance channels. One popular choice for the function $\phi(\cdot)$ is a weighted sum,

$$d(x_j) = d^Y(x_j) + \alpha * d^{Cb}(x_j) + \beta * d^{Cr}(x_j), \tag{8}$$

since for this definition of the block distortion, the frame distortion is also the weighted sum of the luminance and chrominance frame distortions. Clearly the selection of an appropriate α and β has to be done experimentally. Note that by defining the block distortion as above, the inhomogeneous DVF is found optimally, using the information from all three channels. Nevertheless, in the presented results, we set α and β to zero.

6 EXPERIMENTAL RESULTS

As in test model four (TMN4)[11] of H.263 we assign one bit per block to indicate if the motion vector of this block is zero, since this is the most common event. When the motion vector is non-zero, we encode the motion vector difference between the current and the previous block using the entropy table of TMN4. Efficient motion estimation is particularly important for very low bit rate video coding, since the encoding of the DVF can take up to 100% of the available bit rate.

From a theoretical point of view, every possible motion vector of block $b_{l,i}$ should be included in the set $M_{l,i}$, which is the admissible motion vector set for block $b_{l,i}$. This means that for a typical search window of ±15.5 pixels and an accuracy of 0.5 pixel, $|M_{l,i}|$ = 63*63 =3969, which is quite large. Most of these motion vectors, however, are not likely candidates for the optimal path, since they do not correspond well to the real motion in the scene and therefore they lead to a high distortion and a high rate. These motion vectors can be found by performing block matching since they will result in a high matching error. To make the optimization process faster, the prior knowledge about these motion vectors is taken into account. Even though this might be complicated in general, it is easily achieved in the presented framework of DP by reducing the set $M_{l,i}$ of admissible motion vectors of block $b_{l,i}$.

We propose the following scheme for the recursive generation of admissible motion vector sets. An initial motion vector search is conducted for the $2^{n_0} \times 2^{n_0}$ blocks at level n_0 by using block matching with integer accuracy. The K integer motion vectors which lead to the best prediction are kept. Then the set $M_{n_0,i}$ is defined as the set which contains the K integer motion vectors plus their half pixel neighbors. After the set of admissible motion vectors has been defined for the bottom level (n_0), the sets of admissible motion vectors for higher level blocks are defined recursively. A block $b_{l,i}$ only includes a motion vector in $M_{l,i}$ if this motion vector has been selected by all of its child nodes. This leads to the fact that for small blocks, many motion vectors are considered but the bigger the block, the smaller the number of vectors associated with it. This reflects the well known fact that small block sizes lead to small energy in the DFD but not very consistent motion vector fields, whereas bigger blocks lead to consistent vector fields, but the energy in the DFD can be quite high. Our experiments have shown that for $K = 10$ this restriction of the search space does not lead to a performance loss but increases the speed of the algorithm significantly.

We compare the QT-based optimal motion estimation scheme with the fixed block size (16×16) based scheme of TMN4. As pointed out before, we use the same encoding technique for the DVF as TMN4 and hence the difference in performance stems from the optimal tradeoff between the rate necessary to encode the inhomogeneous DVF (QT segmentation and motion vector differences) and the resulting energy in the DFD (the distortion). For the presented experiment, we use the Y-channels of the 176-th and 180-th frames of the QCIF sequence "Mother and Daughter". We select the mean squared error (MSE) as the distortion measure, and employ the peak signal to noise ratio (PSNR) to express its magnitude in dB, (PSNR = $10 * \log_{10}(255^2/\text{MSE})$). We intra code the 176-frame using TMN4 and a quantizer step size of 10. The resulting PSNR for this frame is 33.85 dB . Then the original 180-th frame is used to find the DVF. First we employ the TMN4 block matching scheme which uses the sum of the absolute error and favors the zero motion vector by reducing its error by a constant amount of 100. The resulting bit rate for the DVF and the DFD energy are listed in row "TMN4" in Table 1. The predicted frame and the DVF are displayed in Fig. 9.

Now the QT-based optimal motion estimator is run, where first the maximum rate is set equal to the TMN4 rate, i.e., $R_{max} = 470$ bits. We call this experiment "matched Rate" in Table 1 and the resulting rate and distortion are listed in the corresponding row. Note that for the same rate, the total distortion is reduced significantly, or in other words, a better prediction is achieved. Besides outperforming the TMN4 motion estimator in the objective sense, the proposed QT-based scheme also outperforms the TMN4 motion estimator subjectively. This is due to the inhomogeneous representation of the motion by means of the QT, which enables the QT-based motion estimator to spend more bits in areas with complex motion, i.e., small block sizes are used, and fewer bits in areas with simple motion, i.e., large block sizes are used. The better prediction performance is apparent from the resulting predicted frame and the DVF which are displayed in Fig. 10. For example, the prediction of the left eye and the shirt collar of the mother and the right corner of the frame in the background, is clearly better in the proposed approach. Recall that the n_0 level is scanned by the modified Hilbert scan shown in Fig. 5. To generate the scanning path for an arbitrary QT decomposition, the procedure introduced in Sec. 3.1 is used. In Fig. 11 the resulting overall scanning path which corresponds to the QT decomposition displayed in Fig. 10 is shown.

For the next experiment, the DFD energy of the TMN4 run is matched by setting $D_{max} = 1148$ (=30.6 dB PSNR). We call this experiment "matched Dist." in Table 1 and the resulting rate and distortion are listed in the corresponding row. Note that for the same distortion, the bit rate is reduced significantly (about 30%). The resulting predicted frame and the DVF are displayed in Fig. 12. Again, even though the DFD energy is the same as in Fig. 9, the predicted frame is of higher quality. The same explanation applies as before, that is, the implicit DVF smoothing along the Hilbert scan results in a good DVF. In addition, the inhomogeneous structure of the

DVF fits a real DVF better and hence a better representation of the real DVF can be achieved.

7 SUMMARY

In this paper we presented a QT-based motion estimator which finds the QT-segmentation and the DVF jointly and optimally in the rate distortion sense. The inhomogeneous representation of the DVF results in DVF estimates which are close to the real DVF, and in their efficient encoding. We encode the inhomogeneous DVF using a first order DPCM along the scanning path and the QT data structure. We introduced a procedure to create a spatially non-disruptive scanning path for an arbitrary QT-decomposition, which results in an efficient DPCM encoding of the DVF. We formulated the motion estimation problem as a constrained optimization problem, which we solve using the Lagrangian multiplier method and a multilevel dynamic program. We showed that the proposed motion estimator can easily incorporate multichannel information, such as color. Finally, we presented results of the proposed QT-based motion estimator for a QCIF video sequence and compared them with block matching. The results clearly demonstrated that the proposed scheme outperforms block matching significantly in an objective, as well as, in a subjective sense.

8 REFERENCES

[1] H. Nicolas and C. Labit, "Region-based motion estimation using deterministic relaxation schemes for image sequence coding," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. 265–268, 1992.

[2] M. R. Banham, J. C. Brailean, C. L. Chan, and A. K. Katsaggelos, "Low bit rate video coding using robust motion vector regeneration in the decoder," *IEEE Transactions on Image Processing*, vol. 3, pp. 652–665, Sept. 1994.

[3] J. Lee, "Optimal quadtree for variable block size motion estimation," in *Proceedings of the International Conference on Image Processing*, vol. 3, pp. 480–483, Oct. 1995.

[4] G. J. Sullivan and R. L. Baker, "Efficient quadtree coding of images and video," *IEEE Transactions on Image Processing*, vol. 3, pp. 327–331, May 1994.

[5] H. Musmann, M. Hötter, and J. Ostermann, "Object-oriented analysis-synthesis coding of moving images," *Signal Processing: Image Communication*, vol. 1, pp. 117–138, Oct. 1989.

[6] G. M. Schuster and A. K. Katsaggelos, "An optimal lossy segmentation scheme," in *Proceedings of the Conference on Visual Communications and Image Processing*, pp. 1050–1061, SPIE, Mar. 1996.

[7] F. S. Hill, *Computer graphics.* Macmillan Publishing Company, 1990.

[8] H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," *Operations Research*, vol. 11, pp. 399–417, 1963.

[9] G. M. Schuster and A. K. Katsaggelos, "Fast and efficient mode and quantizer selection in the rate distortion sense for H.263," in *Proceedings of the Conference on Visual Communications and Image Processing*, pp. 784–795, SPIE, Mar. 1996.

[10] G. M. Schuster, *A video compression scheme with optimal bit allocation among segmentation, motion and residual error.* PhD thesis, Northwestern University, Evanston, Illinois, USA, June 1996. Department of Electrical Engineering and Computer Science.

[11] Expert's Group on Very Low Bitrate Visual Telephony, *Video Codec Test Model, TMN4 Rev1.* ITU Telecommunication Standardization Sector, Oct. 1994.

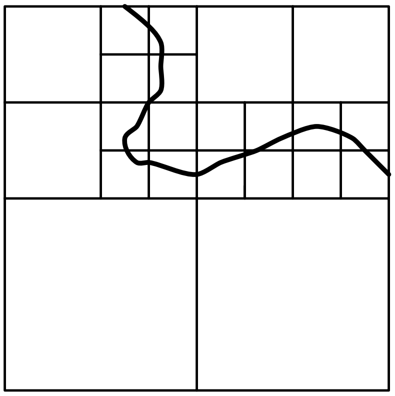

Figure 1: Frame segmented by a quad-tree.

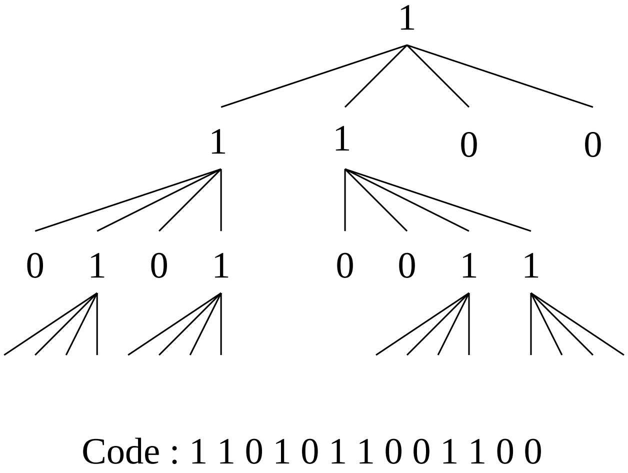

Figure 2: Quad-tree representation of the frame.

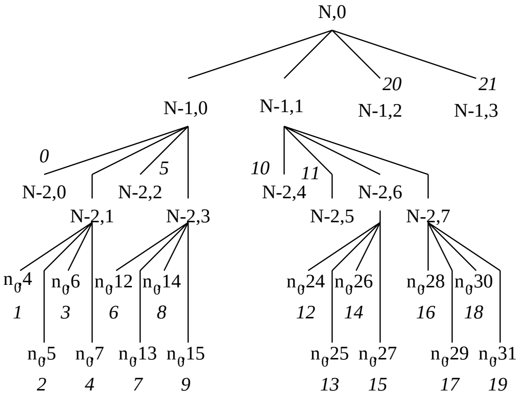

Figure 3: Quad-tree notation. Each node is identified by the ordered pair (l, i), where l is the level and i the number within that level. Also the leafs are numbered from zero to the total number of leafs in the QT minus one.

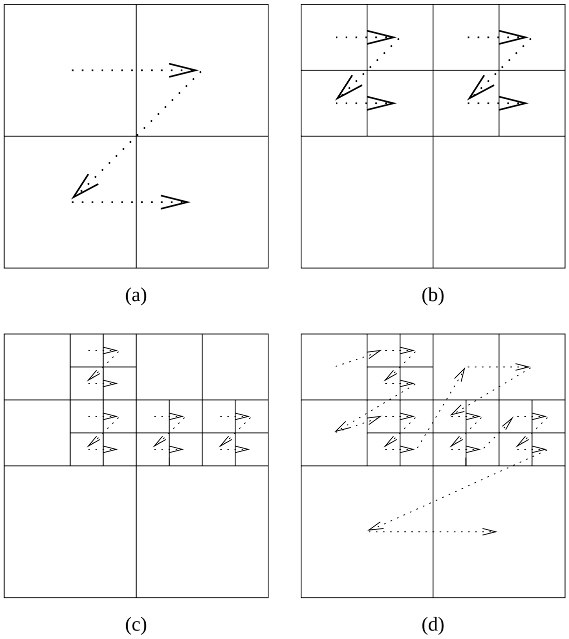

Figure 4: Recursive raster scan for a QT decomposition. Figures (a) through (c) show the raster scan at different levels of the QT and figure (d) shows the resulting overall scanning path.

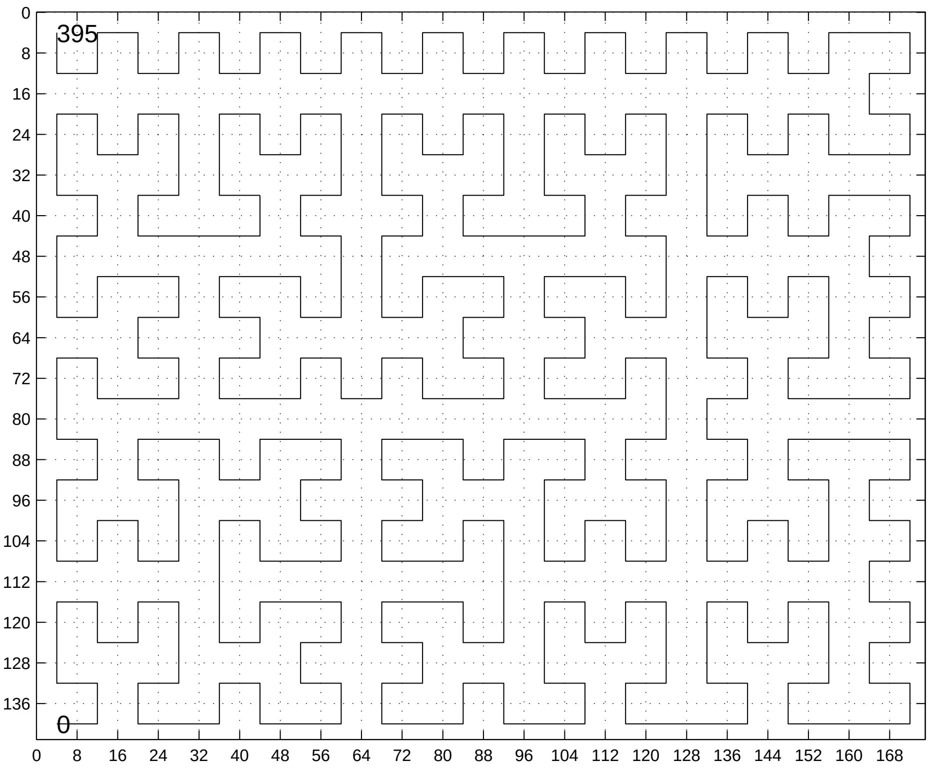

Figure 5: Modified Hilbert scan for level $n_0 = 3$ of a QCIF frame.

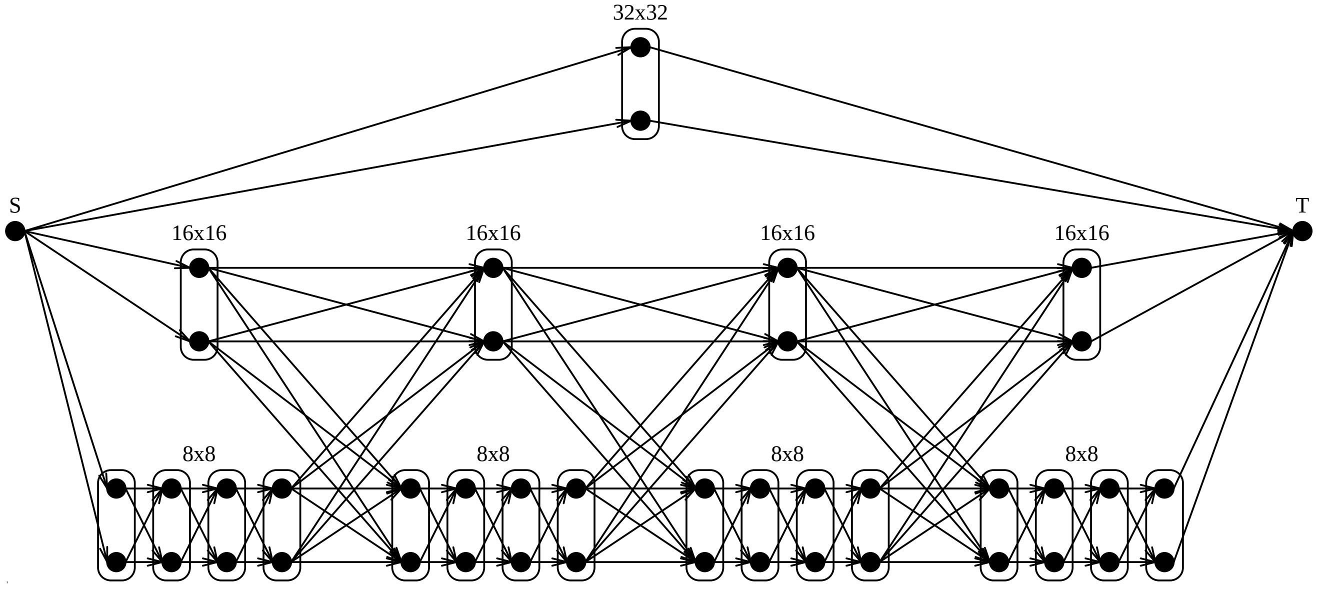

Figure 6: The multilevel trellis for $N = 5$ and $n_0 = 3$.

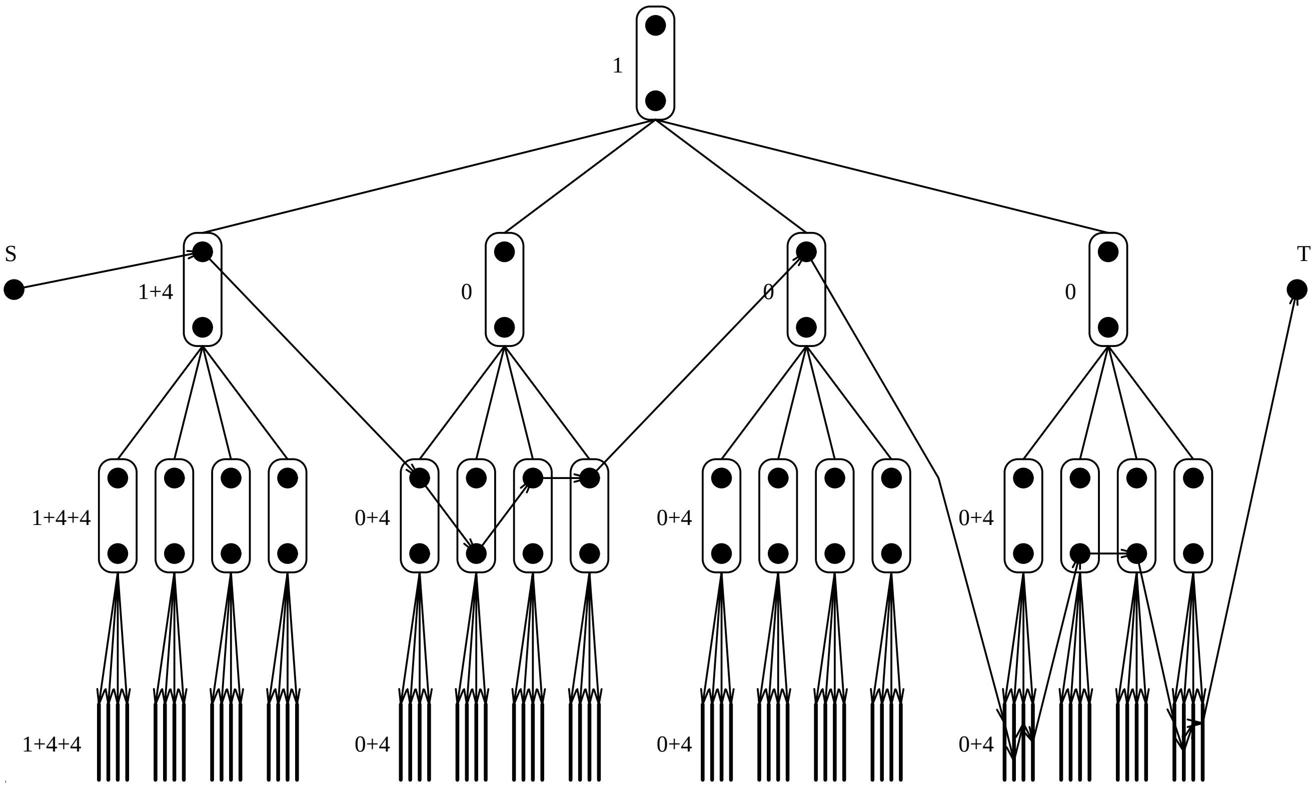

Figure 7: The recursive distribution of the quad-tree encoding cost among the trellis nodes for a quad-tree of depth 4 and an optimal path.

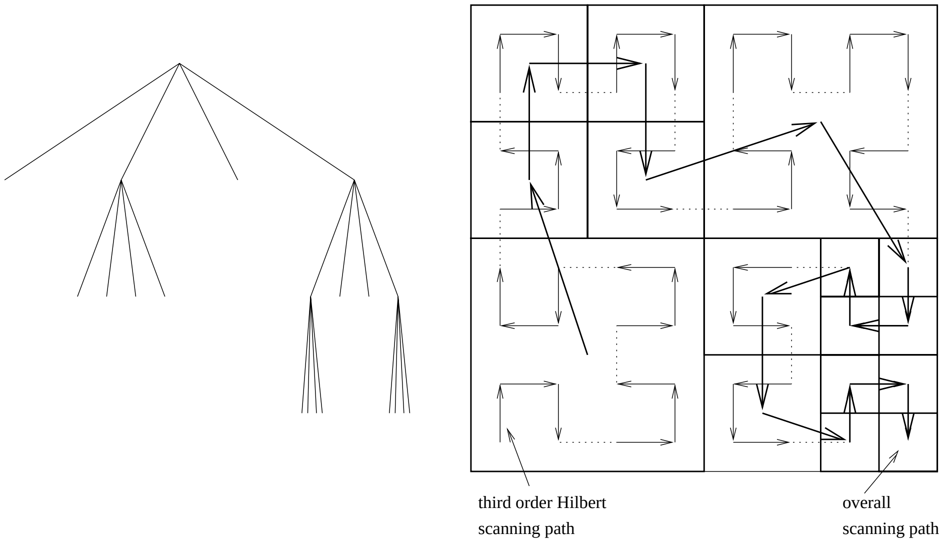

Figure 8: Quad-tree decomposition corresponding to the optimal state sequence. The left figure shows the graphical representation of the QT which corresponds to the frame decomposition in the right figure. The right figure shows the level n_0 scanning path, which is a third order Hilbert curve, and the resulting overall scanning path.

	Rate (bits)	Dist. (PSNR)
TMN4 (Fig. 9)	470	30.6
matched Rate (Fig. 10)	472	31.3
matched Dist (Fig. 12)	344	30.7

Table 1: Comparison between the TMN4 motion estimation algorithm and the proposed optimal motion estimator using the luminance values of the 176-th and 180-th frames of the QCIF sequence "Mother and Daughter".

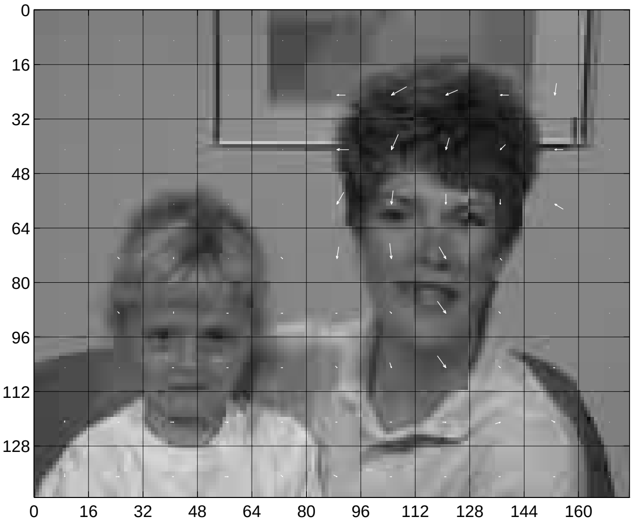

Figure 9: The predicted frame and the DVF for TMN4 block matching.

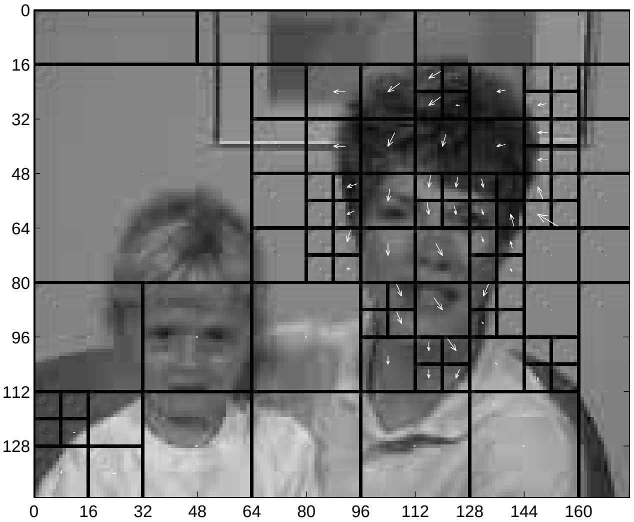

Figure 10: The predicted frame and the DVF for the optimal QT-based motion estimator, when the rate matches the TMN4 rate.

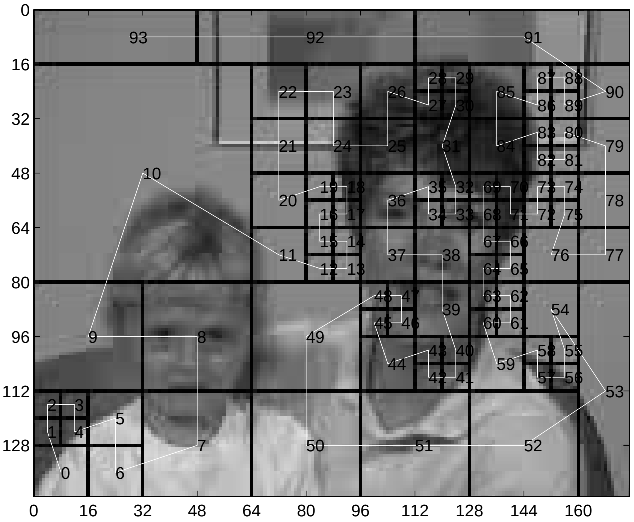

Figure 11: The overall scanning path.

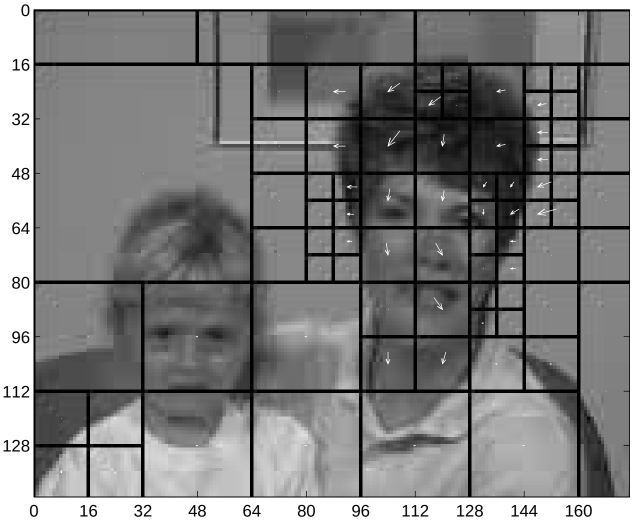

Figure 12: The predicted frame and the DVF for the optimal QT-based motion estimator, when the distortion matches the TMN4 distortion.