RC24748 (W0902-063) February 12, 2009
Electrical Engineering

IBM Research Report

Inter Mode Selection for H.264/AVC Using Time-Efficient Learning-Theoretic Algorithms

Yuri Vatis
Institut für Informationsverarbeitung
Leibniz Universität Hannover
Germany

Ligang Lu, Ashish Jagmohan
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home .

INTER MODE SELECTION FOR H.264/AVC USING TIME-EFFICIENT LEARNING-THEORETIC ALGORITHMS

Yuri Vatis

Institut für Informationsverarbeitung
Leibniz Universität Hannover, Germany

Ligang Lu, Ashish Jagmohan

IBM T. J. Watson Research Center
Yorktown Heights, USA

ABSTRACT

In this paper we present a novel algorithm to speed up the inter mode decision process for the H.264/AVC encoding. The proposed inter mode decision scheme determines the best coding mode between P16x16 and P8x8 based on learning theoretic classification algorithms to discern between mode classes based on the evaluation of a simple set of features extracted from a motion compensated macroblock. We show that the proposed method can reduce the number of macroblocks for P8x8 mode testing by 80% on average, at the cost of only a small loss of 0.1 dB in compression performance.

Index Terms— Mode decision, learning algorithms, H.264/AVC encoder.

1. INTRODUCTION

The latest video coding standard H.264/AVC has provided state of the art tools that can be used to achieve significantly better compression performance than the previous standards over a wide range of bit rates. One of the key tools that H.264/AVC provides for higher coding efficiency is the employment of a large set of coding modes. For instance, in H.264, macroblock prediction can be done in 7 block sizes and intra- and inter-prediction modes. While the addition of coding modes provides greater flexibility to adapt to the signal characteristics for better coding gain, it also increases the encoding computational complexity enormously in searching for the best coding mode. In the H.264/AVC Reference Code [1], the macroblock mode selection is formulated as a rate-distortion optimization process and the Lagrangian minimization technique is used to minimize the cost functional. The computational complexity of the exhaustive search process is very large because it has to encode a macroblock using all possible modes to find the best mode that minimizes the rate-distortion functional. For example, each inter-coding mode needs to go through the entire encoding loop, including the steps of motion estimation, motion compensated prediction, transform, quantization, entropy coding (CABAC), inverse quantization, and inverse transform, to calculate the actual rate and distortion incurred by using this mode. Note that, even with such a high computational complexity method, this mode selection process is still not optimal since it does not take into account the rate-distortion impact of the current macroblock on other blocks which have a direct or indirect prediction dependency on it. Therefore, using a truly optimal mode selection algorithm is unrealistic, and low complexity and effective mode selection algorithms are highly desirable for deploying H.264/AVC in applications.

A significant amount of research efforts has been spent on designing fast mode selection algorithms. Notably, Turaga and Chen [2] presented a pattern classification technique based mode decision scheme using maximum-likelihood (ML) criterion, but their features are not suitable for H.264/AVC mode decision. Jagmohan and Ratakonda [3] developed a supervised binary mode classification scheme for intra/inter mode selection using a learning-theoretic algorithm. More recently, Kim and Kuo [4] proposed an intra/inter mode selection scheme by classifying features derived from spatial correlation, temporal correlation, and motion activity. These classification based algorithms are proposed only for intra/inter mode decision. However the computational complexity from the inter coding mode decision is very demanding because it involves the complete inter coding process, especially the motion estimation step.

In this paper, we will present a new fast inter coding mode selection scheme based on a decision tree classification algorithm. More specifically, our algorithm will analyze a subset of the prediction residues in the frequency domain and derive features that can be quickly calculated and classified by a tree structure. The test results show that our inter coding mode selection algorithm not only is time efficient but also has comparable performance as the H.264/AVC's exhaustive search mode decision algorithm.

In the following, we will first provide a brief overview of the H.264/AVC macroblock coding modes and describe a new decision tree based algorithm for fast inter coding mode selection in Section 2. We will then detail our proposal and describe the feature extraction in Section 3. In Section 4 we will evaluate the performance of our new inter coding mode selection algorithm and compare it with the performance of the state of the art H.264/AVC reference encoder operated in the High-complexity mode. Finally we will summarize our

work and draw conclusions in Section 5.

2. PROPOSED MODE SELECTION FRAMEWORK

The H.264/AVC standard provides seven block sizes to subdivide a macroblock in inter coding mode. The prediction for the current macroblock is generated based on data from prior frames. Operated in the High-complexity mode which provides the best quality at a given bit rate, the encoder determines the best inter mode by minimizing the Lagrangian function [5]:

$$J(s, c, MODE|QP, \lambda_{MODE}) = \quad (1)$$
$$SSD(s, c, MODE|QP) + \lambda_{MODE} \cdot R(s, c, MODE|QP)$$
$$MODE \in \text{[P16x16, P16x8, P8x16 and P8x8 modes]},$$

where SSD is the sum of absolute differences, s is the original signal, c is the reconstructed signal, QP is the given quantization parameter and λ_{MODE} is the Lagrangian factor. The best inter mode is determined by performing motion estimation for all modes. Similarly, the best reference frame for a particular macroblock or a subblock is determined. Finally, the best inter mode is compared with the SKIP mode and the best intra mode. This results in the best coding efficiency but at the cost of the highest computational complexity, which is often prohibitive especially in real-time applications. Reducing the computational complexity is often done by using a subset of the available inter modes, by limiting the number of searches in the available modes based on greedy algorithms, and by approximating the rate-distortion optimization process with a less complex process.

The key idea underlying the proposed mode selection approach is to develop simple heuristics such that macroblocks can be effectively classified into classes of different coding modes, for examples, predictive 16x16 (P16x16) mode and predictive 8x8 (P8x8) mode. Towards this end, we propose the use of learning-theoretic approaches that operate on a set of simple macroblock features to classify the modes. In particular, our algorithm only needs to extract a set of three transform domain features from a predicted macroblock for the decision to select a certain mode. In this paper, we consider P16x16 and P8x8 coding modes as the showcase of describing our mode selection method, for simplicity, and also for the following reasons. Firstly, our primary goal is to develop an efficient but also real-time capable encoder. These two modes are a popular illustrative subset of all modes and many real-time encoders, e.g. the IBM H.264/AVC encoder, often use these two modes. Secondly, the learning-theoretic approaches are most suitable for binary decisions. However, the method and the classifier trees can be readily extended to include other subsets based on the mode decision. The decision between intra and inter modes is not regarded in this paper and the reader is referred to [3].

The proposed algorithm can be described in the following steps:

1. Perform the motion compensated prediction for the macroblock to be coded in P16x16 mode.
2. Based on the extracted features, either terminate here or go to the next step.
3. Perform the motion compensated prediction for the macroblock to be coded in P8x8 mode.
4. Find the best mode based on rate-distortion optimization.

Note that for the last step other functions such as SAD or SATD as used in the Low-complexity mode [5] can also be applied, which further reduce the computational complexity. However, the last decision can be done independently of the second step, therefore we set our focus on the feature extraction and use rate-distortion optimization technique as well as common motion estimation techniques in order to compare our encoder with the JVT encoder operated in the High-complexity mode.

3. FEATURE EXTRACTION

For a given macroblock, the selection between P16x16 and P8x8 modes is made by employing supervised binary classification [6] using a set of transform domain features of the P16x16 motion compensated macroblock.

Consider a data set $\{x_i\}_{i=1}^n$, such that each element x_i belongs to one of two classes C_1, C_2. The aim of binary classification is to infer the class to which each x_i belongs, on the basis of a set of extracted features $\{f_j(x_i)\}_{j=1}^m$. Supervised binary classification does this using two phases - a training phase and a test phase. During the training phase, a set of training data $\{y_k\}_{k=1}^K$ with known class membership labels (C_1 or C_2) is presented to the classifier. The training data and the class labels are used by the classifier to learn a classification function $\Lambda(\cdot)$ such that classification rules

$$y_k \in C_1 \qquad \Lambda\left(\{f_j(y_k)\}_j\right) = 1$$
$$y_k \in C_2 \qquad \Lambda\left(\{f_j(y_k)\}_j\right) = 0$$

minimize the misclassification rate on $\{y_k\}$. The form of the classification function Λ is constrained by the type of classifier used, for example, linear discriminant analysis yields classifier functions which are linear in the feature space. During the test phase, the learned classification function is used to classify the unknown data $\{x_i\}$. The key issues in supervised binary classification are the choice on an appropriate classifier type, and the extraction of a set of features which allows good discrimination between the two classes.

For P16x16/P8x8 discrimination, the motion compensated 16x16 macroblock m is used. The center consisting of 8x8 pels of the current macroblock m is denoted as m_8. Denoting further the 8x8 Hadamard transform of m_8 as $T_H m_8$, the following set of features was found to be effective:

1. The macroblock horizontal, vertical and low two-dimensional frequency content:

$$f_1 = \sum_{j=1}^{7} |T_H m_8(0,j)| + \sum_{i=1}^{7} |T_H m_8(i,0)| + \sum_{i=1}^{3}\sum_{j=1}^{3} |T_H m_8(i,j)|.$$

2. The macroblock low horizontal high vertical and low vertical high horizontal frequency content:

$$f_2 = f_1 + \sum_{i=1}^{3}\sum_{j=4}^{7} |T_H m_8(i,j)| + \sum_{i=4}^{7}\sum_{j=1}^{3} |T_H m_8(i,j)|.$$

3. The macroblock high frequency content:

$$f_3 = \sum_{i=4}^{7}\sum_{j=4}^{7} |T_H m_8(i,j)|.$$

These features are also presented in Figure 1. Intuitively, the P16x16 mode performs better when no high frequencies are contained in the motion compensated macroblock. The presence of high frequencies in the motion compensated macroblock indicates that different parts of this macroblock move differently and therefore P8x8 mode should be tested. In order to keep the complexity of our algorithm low, the 8x8 Hadamard transform is only applied to the center of the motion compensated macroblock. The same type of the classifier trees as in [7] was used.

During the training phase, macroblocks from a large number of video sequences were used as a training set, and the P16x16/P8x8 mode decisions selected by the rate-distortion optimized reference H.264/AVC encoder were used as the known class labels. The classifier tree was trained on this data using the tree-partitioning procedure described in [7]. The training was performed for all quantization parameters in the range from 18 to 45, resulting in a reasonable set covering from the visually lossless quality at one end to where no P8x8 is used due to less efficiency in the rate-distortion sense at the other end. Roughly, the structure of the classifier tree could be subdivided into four types, where the importance of the feature f_3 decreased with the increase of the quantization parameter value. The decision thresholds were found to vary

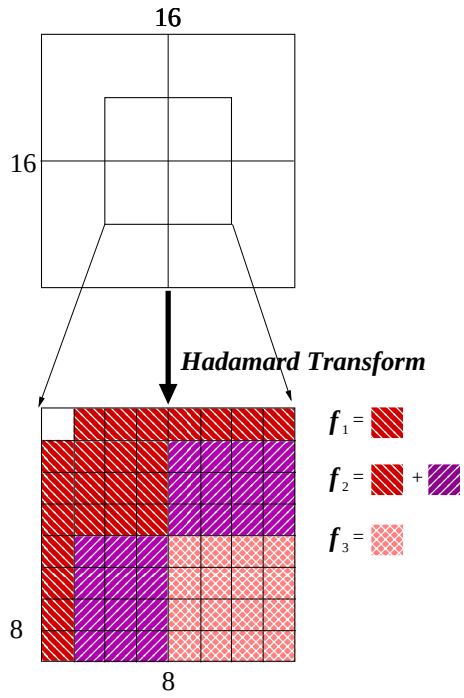

Fig. 1. Decision features.

with changing quantization parameter. One such exemple tree is presented in Figure 2. Note that the P16x16 output implies that the P8x8 mode is not tested. The P8x8 output implies that the P8x8 mode is tested and then the P16x16/P8x8 decision is made on the basis of rate-distortion cost.

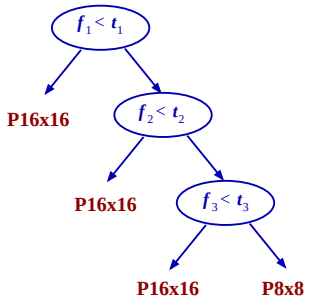

Fig. 2. Example of a learned classifier tree.

For encoding a macroblock from a new video sequence, the feature set $\{f_j\}_{j=1}^{3}$ described above is extracted and the learned classifier tree is used to infer the P16x16-P8x8 decision. Since the learning is performed completely off-line, the proposed mode selection algorithm requires minimal complexity for making the mode selection.

4. EXPERIMENTAL RESULTS

To evaluate the coding efficiency of the new decision tree based fast inter coding mode selection algorithm, we used

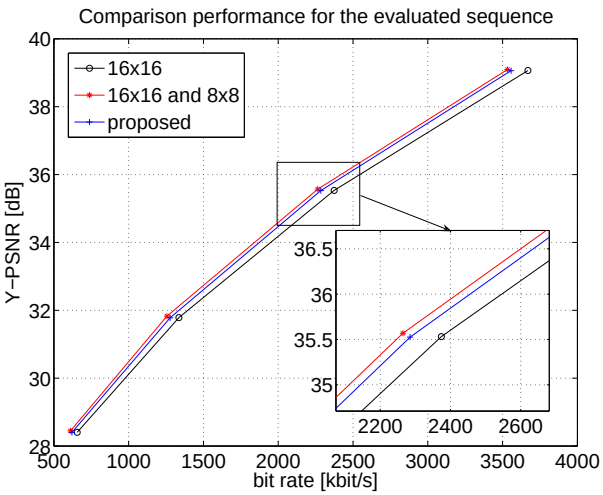

Fig. 3. Compression performance for proposed encoder, R-D optimized P16x16-P8x8 encoder, and P16x16 encoder.

a composite sequence consisting of 2583 frames of the following ten CIF sequences: *Formula1, Flowergarden, Mobile&Calendar, Concrete, Husky, Basketball, TableTennis, Stefan, Tempete* and *Bus*. All simulations were performed using the Baseline Profile of H.264/AVC operated in the High-complexity mode. In Figure 3, three operational rate-distortion curves are presented, showing the compession performance of the reference encoder with enabled intra 16x16, intra 4x4, SKIP and P16x16 modes, the performance of the reference encoder enhanced with additional P8x8 mode and the performance of our encoder. As can be seen, the performance gap between the reference encoder with P8x8 mode and our encoder is approximately 0.1 dB while the performance gap between the reference encoder with P8x8 and without P8x8 mode is 0.4 dB at same bit rates.

QP	8x8 tested	8x8 chosen
20	20.4%	10.9%
24	28.2%	12.3%
28	18.8%	10.1%
32	13.5%	8.0%
36	18.5%	7.0%
40	6.5%	3.5%
44	2.6%	1.3%

Table 1. Number of tested and chosen macroblocks in P8x8 mode.

Table 1 evaluates the reduction of computational complexity compared to the reference encoder with enabled P8x8 mode. For a representative subset of quantization parameters, the number of tested P8x8 modes as well as selected P8x8 modes is given in columns 2 and 3, respectively. In the range from 20 to 36, the average number of macroblocks, where P8x8 mode was additionally evaluated, is approximately 20% for our proposal compared to 100% for the reference encoder. The number of chosen P8x8 macroblocks is 10% on average compared to approximately 30% for the reference encoder. Obviously, 10% of the chosen macroblocks could cover almost all of the coding gain, which shows that our fast inter mode selection algorithm is very effective and efficient.

5. SUMMARY

In this paper, we proposed a novel algorithm for fast inter mode selection for P16x16 and P8x8 modes based on a decision tree classification algorithm. The extracted features are derived from a motion compensated P16x16 macroblock transformed into the frequency domain. The P8x8 mode for a macroblock is tested only in the case that the amount of the high frequencies is higher than a particular threshold, indicating that different parts of the macroblock move differently. Experimental results show that our proposed algorithm only needs one fifth of the complexity of the reference encoder for P8x8 mode testing while capturing 80% of its coding efficiency.

6. REFERENCES

[1] "H.264/AVC reference software version JM14.0, available at http://iphome.hhi.de/suehring/tml/download/old_jm/," July 2008.

[2] D. Turage and T. Chen, "Classification based mode decisions for video over network," *IEEE Trans. On Multimedia*, vol. 3, no. 1, pp. 41–52, March 2001.

[3] A. Jagmohan and K. Ratakonda, "Time-efficient learning theoretic algorithms for H.264 mode selection," in *Proc. IEEE Int. Conference on Image Processing (ICIP)*, Singapore, October 2004, pp. 749 – 752.

[4] C. Kim and C.-C. J. Kuo, "Feature-based intra-/inter coding mode selection for H.264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 4, pp. 441 – 453, April 2007.

[5] K.-P. Lim, G. Sullivan, and T. Wiegand, "Text description of joint model reference encoding methods and decoding concealment methods," in *JVT-O097, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG*, Busan, Korea, April 2005.

[6] R. O. Duda, P. E. Hart, and D. G. Stock, *Pattern classification*, Wiley, 2001.

[7] P. Chou, "Optimal partitioning for classification and regression trees," *IEEE Trans. Pat. Anal. Mach. Intel.*, vol. 13, no. 4, pp. 340–354, April 1991.