# FAST MULTI-REFERENCE MOTION ESTIMATION VIA STATISTICAL LEARNING FOR H.264/AVC

*Chen-Kuo Chiang and Shang-Hong Lai*

Department of Computer Science, National Tsing Hua University,
Hsinchu 300, Taiwan, R.O.C.
{ckchiang, lai}@cs.nthu.edu.tw

## ABSTRACT

In the H.264/AVC coding standard, motion estimation (ME) is allowed to use multiple reference frames to make full use of reducing temporal redundancy in a video sequence. Although it can further reduce the motion compensation errors, it introduces tremendous computational complexity as well. In this paper, we propose a statistical learning approach to reduce the computation involved in the multi-reference motion estimation. Some representative features are extracted in advance to build a learning model. Then, an off-line pre-classification approach is used to determine the best reference frame number according to the run-time features. It turns out that motion estimation will be performed only on the necessary reference frames based on the learning model. Experimental results show that the computation complexity is about three times faster than the conventional fast ME algorithm while the video quality degradation is negligible.

***Index Terms***—Motion estimation, multiple reference frames, H.264, statistical learning

## 1. INTRODUCTION

H.264/AVC, the latest video coding standard of Joint Video Team (JVT), outperforms previous standards, such as MPEG-4 and H.263, in terms of coding efficiency and video quality. This is due to the fact that many new techniques are adopted in this standard, such as variable block size motion compensation, multiple reference frames, directional spatial intra prediction, in-loop deblocking filtering and content-adaptive entropy coding, etc. However, these coding tools also introduce additional computational complexity. Therefore, how to reduce these coding overhead while video perceptual quality can be maintained becomes an interesting issue for H.264 coding system.

Using multiple reference frames can fully make use of temporal correlation of video sequences to achieve high video coding quality. However, not every reference frame is useful for motion estimation. Thus, it turns out to be a reference frame selection problem to choose effective number of reference frames. Recently, many algorithms have been proposed for the multi-reference motion estimation problem. These methods can be classified into two categories. The first one is the rule-based approach. The rules were made according to the criteria to eliminate unnecessary reference frames. In [1], inter SATD, intra SATD and MV compactness were examined to decide whether it is necessary to search more frames. Temporal and spatial content information are checked in the macroblock (MB) levels in [2] to speed up the search process. Kuo and Lu [3] proposed to select suitable reference frames according to the initial search results of an 8x8 size block and only the qualified frames will be further tested in motion estimation. The second category is a semi-statistical learning approach that decides an appropriate number of reference frames based on statistical analysis of a collection of training data. Wu and Xiao [4] employed the statistical distribution of the reference frames along with some rules to determine the optimal reference frame number.

In this paper, we propose a fully statistical learning approach to decide the best reference frame number. The problem of reference frame selection is modeled as a classification problem. First, representative features are chosen according to the analysis from a large number of video sequences. Then, these features are used to train SVM models for classification. Last, an off-line pre-classification approach is employed to decide the best reference frame number according to the run-time features. To our knowledge, this is the first work that introduces a complete machine learning approach to the multi-reference motion estimation problem.

The rest of this paper is organized as follows. Section 2 introduces the extraction of representative features for determining the reference frames. Section 3 presents how a multi-reference motion estimation problem can be formulated as a classification problem and how the SVM classifier is applied here for this problem. Experimental results are shown in Section 4. Finally, Section 5 concludes this paper.

## 2. ANALYSIS AND EXTRACTION OF FEATURES

In this section, we describe the representative features for the selection of multiple reference frames. These features are used to predict the optimal number of reference frames for the current MB. All of them are well chosen and examined carefully in our experiments. The results show that they are helpful to the multi-reference motion estimation.

### 2.1. Analysis for Multi-Reference Motion Estimation

Multi-reference motion estimation (MRME) can achieve higher coding efficiency, but not all the sequences can benefit from MRME. The reasons were investigated in several previous literatures [1] [5]. In summary, searching more reference frames is helpful when

1. A smaller block partition is chosen as the best mode in the variable-block-size motion estimation.
2. It is likely that occlusion or uncovering occurs.
3. The MB is across object boundaries.
4. The MB contains very complicated texture.

### 2.2. Feature Selection

According to the above discussion, we choose the following representative features for each MB:

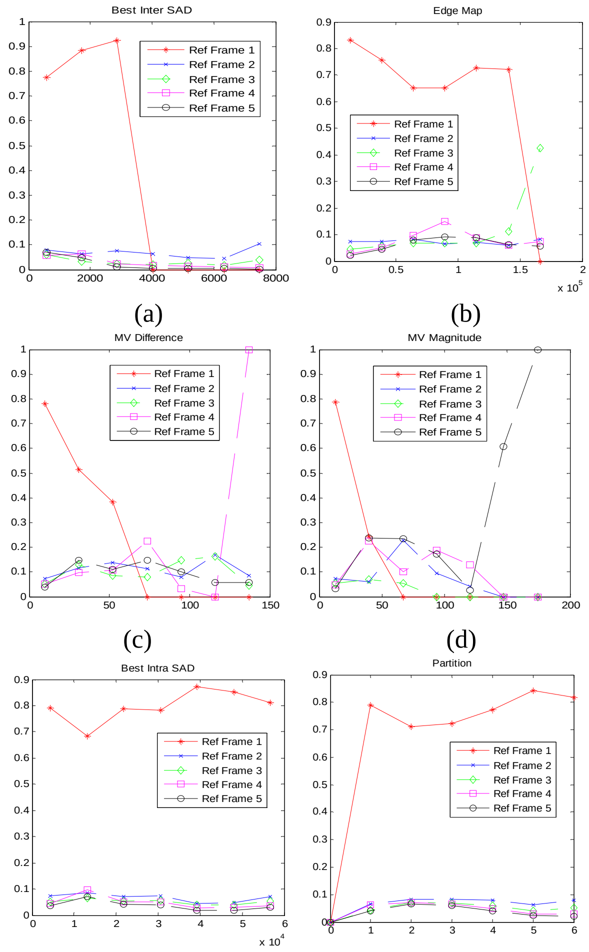
*A. Block Partition:* In the process of motion estimation, MBs are partitioned into different block sizes from 16x16 to 4x4. The size of block partition is labeled from 1 to 7 for feature representation.

*B. Best Inter-SAD:* The lower the best inter-SAD is, the higher probability the current MB contains still background. A background MB has less chance to contain occlusion or uncovering areas. Thus, lower inter-SAD value indicates higher probability of using only one reference frame.

*C. Motion Vector Difference (MVD) and Motion Vector Magnitude (MVM):* MVD may represent the motion smoothness between the current MB and its adjacent MBs. If MVD is small, MBs have similar motion and it is unlikely to cross object boundaries. The MBs with lower MVM values indicates that they are less likely to cross object boundaries as well.

*D. Best Intra-SAD and Gradient Magnitude.* The best intra-SAD is the minimum SAD value after intra prediction of an MB. An MB with a large SAD value after intra prediction usually contains complicated texture or object boundaries. The gradient magnitude of the current MB is defined as the summation of the gradient magnitudes of all pixels inside the MB obtained by applying the Sobel operator. The gradient magnitude can also reflect whether the texture is strong or not in this MB.

Fig. 1 shows the probability of reference frame 1~5 under different features. Fig. 1(a)&(b) indicate the dropping of the probability of reference frame 1 as the amount of the specified features increase to some levels. Fig. 1(c)&(d) show that the decreasing probability of reference frame 1 and the increasing probability of reference 4 and 5 as the amount of features increases. In Fig. 1(e)&(f), the probability of reference 1 is rather high in all conditions. Therefore, the best intra-SAD and Block Partition features may not be so effective. However, they are still helpful in the combined features in the SVM classifier.



(a) (b)

(c) (d)

(e) (f)

**Fig. 1**. The probability of reference frame 1~5 with respect to (a) best inter-SAD, (b) gradient magnitude, (c) MVD, (d) MVM, (e) best intra-SAD and (f) block partition.

## 3. FAST MULTI-REFERENCE MOTION ESTIMATION VIA STATISTICAL LEARNING

We formulate the multi-reference motion estimation problem as a classification problem and provide a solution in this section. Firstly, ME is performed on the first reference frame. Then, the number of necessary frames is predicted based on the above representative features.

### 3.1. Problem Formulation of Reference Frame Selection

In H.264 reference software, ME can be use up to 16 reference frames. For coding efficiency, the number is

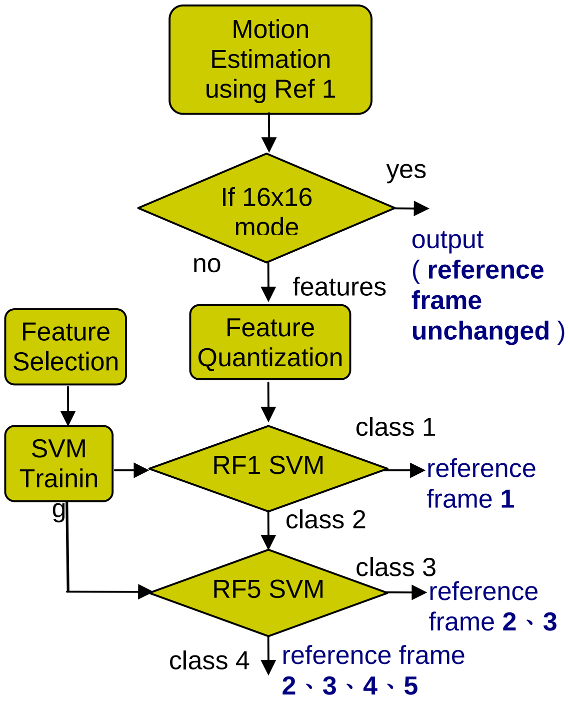


**Fig. 2**. Flow chart of the proposed reference frame prediction algorithm for motion estimation.

usually set to 5. Then, for each MB, we can define five classes, using one, two or up to five reference frames, respectively. Since we can obtain the probability distribution of the selected number of reference frames from a collection of training data, as shown in Fig.1, the experiments indicate that the class reference 2, 3, 4, 5 usually have more similar distributions than those with class reference 1. Thus, we define two binary classifiers: RF1 and RF5 SVM classifiers. The flow chart of the proposed reference frame prediction algorithm by using these two SVM classifiers are shown in Fig. 2. The MBs are classified into four classes. Positive data out of RF1 SVM belongs to class one (C1). Otherwise, it belongs to class two (C2). Similarly, the RF5 SVM further divides C2 output into class three (C3) and class four (C4). The six features, described in subsection 2.2 are used for the SVM classification, including Best Inter/Intra SADs, Motion Vector Difference (MVD), Motion Vector Magnitude ($MV_{mag}$), Block Partition (BP) and Gradient Magnitude (GM). Thus, the binary classification can be decided from the class conditional probabilities given these features.

### 3.2. Training and Pre-Classification

The decision rules can be defined for our classification problem based on class conditional probabilities. However, it is difficult to model the joint probability of those features from limited training samples. The Support Vector Machine (SVM) [6] is employed here for this classification problem.

The training data is obtained by applying H.264 reference code JM 11.0 to three video sequences; namely, News, Container and Coastguard videos. Then, the aforementioned features and the number of the best reference frames determined from the multi-reference ME (ground truth) are collected as the input training samples for SVM. Experimental results show that the accuracies of cross validation for the RF1 and RF5 SVM classifiers are 85.58% and 95.99% respectively, which show high classification rate by using SVM on this problem.

For the consideration of real-time encoding, it takes too much time for run-time classification for SVM. Thus, an off-line pre-classification approach is proposed to minimize the computation time involved in the classification procedure. The idea is to generate all possible combinations of the feature vectors. Then, pre-classify them with SVM. To reduce the large number of possible data combinations, Lloyd-Max quantizer [7] is introduced to quantize each feature based on its feature distribution determined from the training data. It has adaptive step size which provides better approximation of a distribution than the uniform quantizer.

By using the trained SVM classifiers, we can decide the class for all possible input samples. The classification results are stored. During the encoding, it only needs to collect the necessary features, quantize them and search the look-up table for the corresponding SVM classification result. Hence, the computation time in the classification is significantly reduced by using this off-line pre-classification approach.

### 3.3. Proposed Multi-Reference ME Algorithm

The proposed multi-reference ME algorithm is performed on each MB, as Fig. 2. The procedure is given as follows:

Step 1) Perform ME on reference frame 1.
Step 2) If 16x16 is chosen as the best mode, using the previous reference frame number for ME. Go to Step 9.
Step 3) Collect six features.
Step 4) Quantize features with the Lloyd-Max quantizer
Step 5) Apply RF1 SVM classifier via table look-up. If the result is C1, go to Step 9.
Step 6) Apply RF5 SVM classifier via table look-up. If the result is C3, go to Step 7. Otherwise, Step 8
Step 7) Perform ME on reference 2 and 3. Go to Step 9
Step 8) Perform ME on reference 2 to 5. Go to Step 9
Step 9) Go to Step 1 and proceed to the next MB.

Notice that according to the analysis in [1], reference frames are usually unchanged when the 16x16 mode is selected. Thus, this is set as an early termination criterion in Step 2. In the proposed algorithm, the features are quantized

**Table 1**. Average speedup ratio of ME time.

| Sequence | ME Time (s) | | Speedup |
|---|---|---|---|
| (QP 28) | EPZS | Proposed | ratio |
| HallMonitor | 158.97 | 66.19 | 2.40 |
| M_D | 180.68 | 67.10 | 2.69 |
| Stefan | 316.52 | 96.14 | 3.29 |
| Akiyo | 144.44 | 61.78 | 2.34 |
| News | 180.84 | 71.14 | 2.54 |
| Coastguard | 339.99 | 95.33 | 3.57 |
| Average | | | 2.81 |

**Table 2**. PSNR decrease comparison

| Sequence | PSNR (dB) | | Decrease |
|---|---|---|---|
| (QP 28) | EPZS | Proposed | Diff |
| HallMonitor | 37.64 | 37.64 | 0.00 |
| M_D | 38.89 | 38.85 | 0.04 |
| Stefan | 35.26 | 35.14 | 0.12 |
| Akiyo | 39.88 | 39.86 | 0.02 |
| News | 38.2 | 38.18 | 0.02 |
| Coastguard | 34.66 | 34.61 | 0.05 |
| Average | | | 0.042 |

**Table 3**. Bitrate increase comparison

| Sequence | Bitrate | | Increase |
|---|---|---|---|
| (QP 28) | EPZS | Proposed | Diff |
| HallMonitor | 391.38 | 406.82 | 0.0395 |
| M_D | 234.59 | 247.28 | 0.0541 |
| Stefan | 1976.93 | 2196.8 | 0.1112 |
| Akiyo | 211.59 | 222.27 | 0.0505 |
| News | 400.98 | 413.85 | 0.0321 |
| Coastguard | 1342.6 | 1419.9 | 0.0576 |
| Average | | | 0.0575 |

and the SVM classification results are obtained simply via table look-up. Thus, they introduce insignificant computational complexity for the encoding process. Compared with the original JM11.0, the extra overhead of the proposed algorithm contains the calculation of gradient magnitude for each MB.

## 4. EXPERIMENTAL RESULTS

The proposed algorithm is implemented in JM11.0 using the fast ME, EPZS. The motion search range is set to 32 and the maximal number of reference frames is set to 5. The RD optimization and the CABAC entropy encoding are enabled in our experiments. All test sequences are in CIF format and tested on an Intel Core2 CPU 6320 at 1.86 GHz. All frames except the fist frame are encoded as P-frames. The QP is set to 28. Since EPZS was proven to have the PSNR and bitrate results similar to those by full search ME, we applied EPZS and the proposed algorithm to six test sequences. Table 1 shows the speedup ratios of the results for different sequences. The speedup ratios are calculated in terms of the total ME execution time. It indicates that the execution time is nearly three times faster than the fast ME algorithm, i.e. EPZS in JM11.0. The PSNR decrease and bitrate increase comparison are listed in Table 2 and Table 3, respectively.

From these results, it is obvious that our algorithm has similar coding quality compared with EPZS. Both the average PSNR and bitrate deviations are negligible. However, the ME computational efficiency of our proposed algorithm are significantly superior to that of EPZS.

## 5. CONCLUSION

In this paper, we presented a multi-reference motion estimation algorithm based on statistical learning. In this work, several representative features are employed to decide the best reference frame number. The feature analysis shows that they can provide good discriminating features for this problem. To our knowledge, this is the first work that introduces a statistical learning model and provides a complete framework for the multi-reference motion estimation problem. Experimental results show that the execution speed of our algorithm is about three times faster than that of the existing fast ME method while achieving nearly the same compression quality in terms of PSNR and bitrate. In the future work, we would like to investigate more reliable features to improve classification rate. On the other hand, in our experiments, we used median-motion sequences as our training samples. A variety of videos of different motion patterns, such as fast, median and slow motions, can be included into the training data.

## ACKNOWLEDGEMENTS

This work was supported in part by the National Science Council, Taiwan, R.O.C., under grant 97-2220-E-007-007.

## REFERENCES

[1] Y.-W. Huang, et al., “Analysis and reduction of reference frames for motion estimation in MPEG-4 AVC/JVT/H.264,” in Proc. IEEE ICASSP, Apr. 2003.

[2] Q. Sun, X. H. Chen, X. Wu, and L. Yu, “A content-adaptive fast multiple reference frames motion estimation in H.264,” in Proc. IEEE ISCAS, pp. 3651-3654, May 2007.

[3] T.-Y. Kuo and H.-J. Lu, “Efficient reference frame selector for H.264,” IEEE Trans. on Circuits & Systems for Video Technology, vol. 18, no. 3, March 2008.

[4] P. Wu, C.-B. Xiao, “An adaptive fast multiple reference frames selection algorithm for H.264/AVC,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Apr. 2008

[5] Y. P. Su and M. -T. Sun, “Fast multiple reference frame motion estimation for H.264/AVC,” IEEE Trans. on Circuits & Systems for Video Technology, vol. 16, pp. 447–452, Mar 2006

[6] Corinna Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, pp.273-297, 1995.

[7] T. M. Cover, Information Theory, Wiley-Interscience, 1991