# FEATURE-MATCHING BASED MOTION PREDICTION FOR HIGH EFFICIENCY VIDEO CODING IN CLOUD

*Xiang Zhang, Shiqi Wang, Shanshe Wang, Siwei Ma, and Wen Gao*

Institute of Digital Media, Peking University, Beijing 100871, China

## ABSTRACT

Visual features of images and video frames have become pervasive and maturely developed in extensive research fields such as computer vision and visual search. For realtime retrieval applications, the compact visual features should be transmitted and stored at server side in cloud. These local feature descriptors are characterized by the invariance properties for the variances caused by camera motion, illumination changing, occlusion and different viewpoints. Inspired by these properties, the typical scale-invariant feature transform (SIFT) descriptor is leveraged to improve the video coding efficiency in this work. In particular the predicted motion using SIFT matching is used for merge mode and motion vector prediction (MVP) in the high efficiency video coding (HEVC) standard. A hierarchical motion derivation framework aiming at achieving robust and effective MVP is further proposed. Experimental results have shown that the proposed method can efficiently improve the coding performance according to the accurate feature-matching.

***Index Terms***— SIFT, merge mode, motion vector prediction, high efficiency video coding.

## 1. INTRODUCTION

With the marvellously increasing number of visual contents in the cloud environment as well as the rapidly developing performance of smart mobile terminals, the applications of mobile search that utilize the image/video visual features, such as image/video retrieval [1], copy detection [2], objective detection [3] and super-resolution [4], are dramatically growing up. A typical visual descriptor used for visual retrieval is always composed of two main elements, namely a global descriptor and a group of local descriptors. The global descriptor is an overview of a given image which sketches the most significant characteristics in the image. However global descriptor loses all the details as well as location information and fails in object matching and localization, motivating the usage of local descriptor to address this issue. Besides local features have the advantage of the invariance property for the variances caused by camera motion, illumination changing, occlusion and different viewpoints. Many effective and robust local descriptors have been studied for decades, such as *Scale-Invariant Feature Transform (SIFT)* [5] and Speeded Up Robust Features (SURF) [6].

The feature based image compression is proposed in [7], where the SIFT descriptors and the down-sampled image are compressed and the decoder reconstructs the image using similar images in cloud. More recently, the feature is utilized in high efficient image set coding [8]. Consequently the coding performance is highly dependent on other images rather than the compressed image. In reality, the inter-frame correlation are much stronger than the inter-picture correlation. In [9, 10], the feature based motion compensation is introduced to perform fast motion estimation in video coding. However in current video coding design, the traditional block based matching is still employed due to its good tradeoff between efficiency and accuracy [11, 12]. In the state-of-the-art high efficiency video coding (HEVC), the coding block can be directly partitioned into multiple small sub-blocks in a quadtree structure. This quadtree splitting process can be iteratively performed until the size of a sub-block reaches the minimum allowed size, which is always $8 \times 8$ in the common test condition.

In this paper, we leverage the compact features stored in cloud to provide a more accurate motion predictor for video coding. The famous SIFT feature is employed in this work because it is robust and efficient for geometrical transform, illuminance change and general motion. The typical feature extraction process consists of two stages, i.e. the interest point detection and the descriptor calculation. The extracted feature has threefold benefits for interpicture prediction in current hybrid video coding context:

- Providing more accurate candidates for merge mode. It is effective especially when all neighboring blocks of current block are coded using intrapicture mode or located outside the current slice or tile boundary, because the traditional merge candidate is not available in these cases.

- Providing more accurate MVP for non-merge mode. If the motion around current position is not consistent in both spatial and temporal domain, the additional MVP derived from feature matching could be a good substitution and thus helps in improving the coding performance. Especially for the videos with large motion that

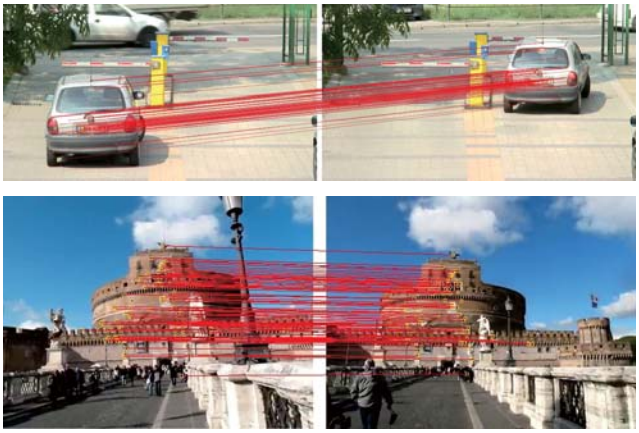

**Fig. 1**. The feature based matching after RANSAC [13] among video frames.

beyond the search range, the feature matching is an effective way in finding the accurate MVP. As illustrated in Fig. 1, the matching is accurate in spite of the geometrical transform and luminance changes.

- Initializing the search origin of motion estimation. Most of the fast motion estimation methods are based on the gradient-descent algorithm, which requires a good origin for faster converging.

To obtain roust and effective motion prediction, we propose a hierarchical motion derivation framework where the MV can be derived from temporally or spatially neighboring blocks. Experimental results on test sequences have shown that the proposed scheme can successfully predict the motion information among video frames and improve the coding performance by -1.13% on average.

The remainder of this paper is organized as follows. The related motion compensation work in current HEVC is introduced in Sec. 2 The details of the proposed method will be discussed in Sec. 3. In Sec. 4 the experimental results are given and analyzed. Finally we conclude this work in Sec. 5.

## 2. RELATED WORK

Motion estimation is critical to video compression due to the strong correlations among video frames. In the state-of-the-art high efficiency video coding (HEVC) standard [14], the video frame is partitioned into quadtree-based coding blocks. For inter-frame prediction, each block can be predicted using a best match block within a limited search range in previous coded frames. It is also called motion estimation (ME) in video coding framework, which is one of the most computationally expensive operations in the encoder. The derived motion information generally consists of the horizontal and vertical displacement vector, namely the motion vector (MV), and one or two reference picture indices.

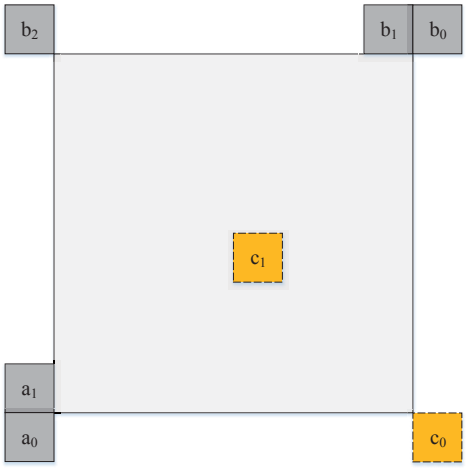


**Fig. 2**. Positions of spatial and temporal candidates of merge mode and advanced mvp (AMVP) where $\{a_1, b_1, b_0, a_0, b_2\}$ are spatial candidates and $\{c_0, c_1\}$ are temporal candidates [16].

HEVC includes a merge mode to derive the motion information from spatially or temporally neighboring blocks [15]. Incorporated with this mode, a merged region containing multiple coding blocks sharing the same motion information is formed. Namely the motion parameters of a merged block can be derived from neighboring block by only transmitting an index indicating which block is referred to.

The set of possible candidates in the merge mode consists of spatial neighboring candidates, a temporal candidate and predefined candidates as shown in Fig. 2. First the spatial candidate positions are checked according to the order $\{a_1, b_1, b_0, a_0, b_2\}$. If the corresponding block located at the specific position is intra-predicted or outside the slice/tile boundary, it is considered as unavailable. The redundant entries where candidates have exactly the same MV are excluded from the candidate list. Then the temporal candidate is chosen from the set $\{c_0, c_1\}$. The size of candidate list $S$ is specified in the slice header. If the number of candidates is large than $S$, only the first $S$ candidates are retained. Otherwise the remaining candidates are generated using the predefined MVs such as zero motion vector for P slices. In HEVC skip mode is a special case of merge mode when all coefficients in current block are equal to zero, where only a skip flag and corresponding merge index should be signaled to decoder.

If a block is not coded using skip or merge mode, the motion vector is compressed differentially using a motion vector predictor (MVP), i.e. the advanced motion vector prediction (AMVP) technique. The candidates of AMVP are also derived from neighboring blocks as in merge mode. However, only a much lower number of candidates is allowed in

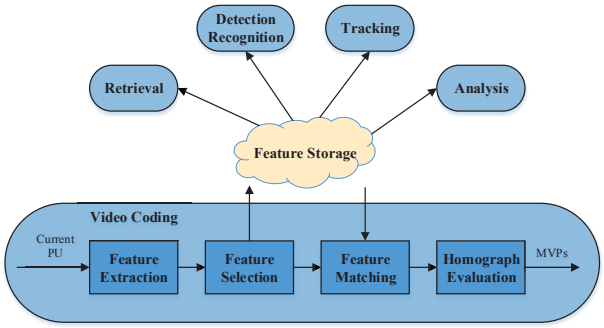


**Fig. 3**. Flowchart of the feature-based motion prediction and extended applications along with this framework.

AMVP. For spatial candidates, the first one is chosen from the set of left positions $\{a_0, a_1\}$ and the other from above positions $\{b_0, b_1, b_2\}$. When the number of candidates is not equal to two after excluding the same MVP, the temporal candidate is included from $\{c_0, c_1\}$. Given the MVP, a motion vector difference or MVD for short can be calculated as the difference between the MVP and the actual searched MV. Generally coding the MVD instead of MV can decrease the compressive cost because the motions of neighboring blocks are highly correlated. Moreover, the derived MVP act as the initialization of the motion search origin for speeding up the convergency process in ME.

## 3. FEATURE-BASED MOTION PREDICTION

The framework of the proposed method is demonstrated in Fig. 3, including feature extraction, selection, matching and homograph evaluation stages. The extracted features after feature selection approach could be compressed and transmitted to cloud server for other feature-guided applications such as retrieval, detection, recognition, tracking and analysis. The details of the feature-based motion prediction will be given this section.

### 3.1. Feature Extraction

Typical feature extraction is composed of two steps, namely interest point detection and descriptor calculation. An interest point is a clear, well-defined, mathematically well-founded position in an image space and can be detected robustly with illuminance variations as well as geometrical changes including translation, rotation, scaling etc. In this work, the interest points are the extreme points detected over a multi-scale Laplacian pyramid in a coarse-to-fine way. This multi-scale structure makes the descriptors scale-invariant that can be detected at different sizes. Then the patch around the interest point should be rotated to its main orientation, which is estimated based on the local gradients. This process guarantees that the descriptor is invariant to rotation. After that the descriptor calculation is performed to generate the 128-dimensional SIFT descriptor vector according to the histogram of gradient orientation within a Gaussian weighted window [5].

### 3.2. Feature Selection

The feature selection approach [17] aims at picking out the descriptors that are most probable as the key point in visual search. The feature selection process has twofold benefits: 1) maximizing the retrieval performance under a constrained bandwidth, and 2) eliminating the useless calculations such as in feature matching and homograph evaluation. In this process, a positive value is assigned to each feature as a "keypoint relevance function" of its characteristics including scale space, orientation and coordinates etc, each characteristic is referred to as a factor of the function. Let $r_{s_i}$ denote the function value (or keypoint relevance) of feature $s_i$, then the features are reordered such that $r_{s_1} \leq r_{s_2} \leq \ldots \leq r_{s_N}$. The first $M$ features with maximum function value are retained for further processing.

Suppose that different factors are mutually independent events, the relevance function can be written as the multiplication of each conditional probabilities,

$$r = \prod_i \hat{P}(c = 1 | f_i \in F_i) \tag{1}$$

where $i$ indicates the factor e.g. scale, orientation, coordinates and respond value. $f_i$ is the value of each factor within the range $F_i$. $c$ is a binary value indicating whether the feature is matched (=1) or mismatched (=0).

The conditional probability can be estimated in the training process as follows,

$$\begin{aligned} \hat{P}(c = 1 | f_i \in F_i) &= \frac{\hat{P}(f_i \in F_i \cap c = 1)}{\hat{p}(f_i \in F_i)} \\ &= \frac{\sum_{n=1}^{N} \kappa(f_i \in F_i) c_n}{\sum_{n=1}^{N} \kappa(f_i \in F_i)} \end{aligned} \tag{2}$$

where the $\kappa$ and the $c_n$ are indicator of membership in $F_i$ and whether matched, both either 0 or 1. $N$ is the size of the training set.

### 3.3. Feature Matching and Homograph Evaluation

The feature matching approach aims at searching the best match in previous frames for each descriptor. Consequently the extracted features in each frame should be catched in memory buffer, and the buffer size is determined by the reference picture set in video coding. The matching evaluation is judged from current frame to previous frame by ratio test as recommended in [5], where the ratio of closest distance to

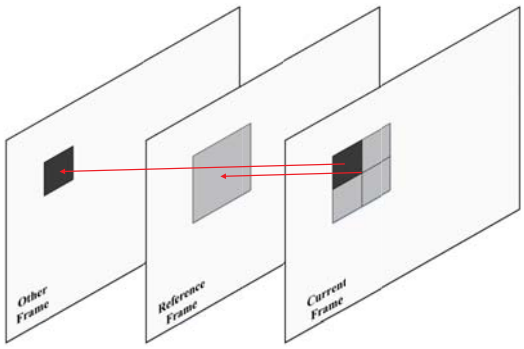


**Fig. 4**. Illustration of hierarchical motion prediction. The motion vector predictor of current block can be derived from temporally neighboring block or inherited from its parent block.

the next closest distance is used as a criterion for distinctiveness to determine the keypoint matches (correspondences) between two images. The distance between two SIFT descriptors is measured by L1 norm. Then some wrong matched pairs are excluded according to the homograph evaluation by the well-known random sample consensus (RANSAC) technique [13], however only two parameters i.e. horizontal and vertical displacements are introduced in the transform matric.

### 3.4. Hierarchical Motion Prediction

The matched feature pairs act as the predictor in interframe reference mode because they can provide all the motion information including motion vector and reference frame index. For maximum utilizing the information, we propose a hierarchical motion prediction scheme to adapt the HEVC coding structure where the motion vector can be derived (or inherited) from temporally and spatially neighboring regions as illustrated in Fig. 4. If current coding block detect no inlier match in reference frame, a scaled motion vector can be derived from other reference frames, where the scale factor is determined by the distances to current frame. Alternatively the motion vector could also be inherited from upper level coding block. For instance, if current block is with size $8 \times 8$, the corresponding motion vector can be predicted by that in its parent block i.e. the $16 \times 16$ block as shown in Fig. 4. In this work the temporal extension is preferred because temporal content is correlated in most cases.

The MVP candidate list is then constructed as illustrated in Fig. 5. The candidates derived from spatially and temporally neighboring blocks are first added to the list after removing the duplicated MVPs as described in Sec. 2. Then if the number of candidates in the list is lower than the specified maximum size $S$, the predicted MVPs using feature matching among video frames performs as additional candidates. Lastly the predefined MVs are added in case the candidate

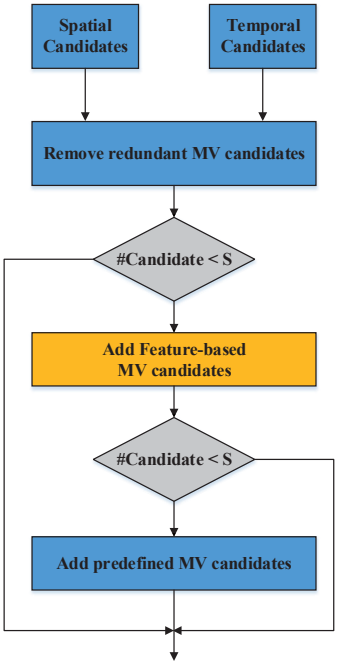


**Fig. 5**. Flowchart of the MV candidate list construction.

list is still not full. If the feature based MVP is employed, the signaling of reference frame and MVP indexes is not required because they can be derived at decoder side in the same way. Besides, the MVP accuracy can be improved compared to the traditional derivation from neighboring blocks. Such that the probability of merge mode is increased and the bits for coding the MV difference (MVD) can be also reduced.

## 4. SIMULATION RESULTS

To evaluate the proposed scheme, we apply it to the state-of-the-art HEVC reference software HM-14.0 using the HEVC Main Profile low-delay configuration. Specifically, the Group of Picture (GoP) size is 4 with only P slices. The test sequences are used in common video coding with different resolutions including 1080P ($1920 \times 1080$), WVGA ($832 \times 480$) and WQVGA ($416 \times 240$).

We first apply only the feature based MVP scheme to the non-merge coding blocks, then compared with that both the feature based MVP and merge mode are performed. BD-rate is employed as the evaluation metric where the negative values indicate the coding gain. The corresponding experimental results are reported in Tab. 1 where it can be concluded that the coding performance is improved for both luminance and chrominance components by leveraging the proposed scheme. The performance of merge mode is further increased comparing with the MVP only mode, indicating the proposed method is efficient for both MVP and merge mode.

**Table 1**. Coding performance improvements using proposed scheme. We compare the MVP only and MVP+merge schemes with the HM-14.0 anchor. BD-rate is employed as the evaluation metric where the negative values indicate the coding gain.

| | MVP only | | | MVP+merge | | |
|---|---|---|---|---|---|---|
| Sequences | Y | U | V | Y | U | V |
| BasketballDrive@1080P | -0.56% | -0.27% | -0.73% | -0.61% | 0.20% | -0.89% |
| BasketballDrill@WVGA | -0.91% | -0.57% | -1.29% | -1.29% | -0.78% | -1.53% |
| RaceHorses@WVGA | -0.82% | -1.32% | -1.16% | -1.04% | -1.16% | -1.71% |
| BasketballPass@WQVGA | -0.57% | -1.62% | -0.70% | -1.12% | -1.56% | -1.39% |
| BQSquare@WQVGA | -1.55% | 0.11% | -1.16% | -1.45% | 0.34% | -0.83% |
| BlowingBubbles@WQVGA | -0.75% | -0.87% | -0.40% | -1.26% | -0.80% | -1.14% |
| RaceHorses@WQVGA | -1.04% | -1.55% | -0.99% | -1.15% | -1.42% | -0.98% |
| Average | -0.88% | -0.87% | -0.92% | -1.13% | -0.74% | -1.21% |

## 5. CONCLUSIONS

In this work, we propose a novel feature based motion prediction framework for high efficiency video coding. Specifically, the extracted SIFT features are utilized to provide motion vector prediction (MVP) for current coding block. This MVP performs as an additional candidate for merge coding blocks or a substitution MVP for non-merge blocks. Furthermore we propose a hierarchical motion derivation framework for achieving robust and effective MVP. Experimental results have shown that the proposed method can improve the coding efficiency because the feature based motion prediction is more precise than that derived from neighbor blocks, especially for sequences with large motions.

## 6. REFERENCES


[1] Pengfei Xu, Lei Zhang, Kuiyuan Yang, and Hongxun Yao, "Nested-SIFT for efficient image matching and retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 34–46, July 2013.

[2] Xiaoguang Gu, Dongming Zhang, Yongdong Zhang, Jintao Li, and Lei Zhang, "A video copy detection algorithm combining local feature's robustness and global feature's speed," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 1508–1512.

[3] Wan-Lei Zhao and Chong-Wah Ngo, "Flip-invariant SIFT for copy and object detection," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 980–991, March 2013.

[4] Huanjing Yue, Jingyu Yang, Xiaoyan Sun, and Feng Wu, "Sift-based image super-resolution," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2013, pp. 2896–2899.

[5] D.G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157 vol.2.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*, pp. 404–417. Springer, 2006.

[7] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu, "Cloud-based image coding for mobile devices–toward thousands to one compression," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 845–857, June 2013.

[8] Xinfeng Zhang, Yabin Zhang, Weisi Lin, Siwei Ma, and Wen Gao, "An inter-image redundancy measure for image set compression," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2015.

[9] Tingrong Zhao and T. Ohtsuki, "A fast motion-compensation scheme for video coding using feature vector matching," in *IEEE Asia-Pacific Conference on Circuits and Systems*, Nov 1998, pp. 635–638.

[10] Xiao chun Zou, Ming yi He, Xin bo Zhao, and Yan Feng, "A robust feature-based camera motion estimation method," in *Asia-Pacific Conf on Innovative Computing Communication and Information Technology Ocean Engineering (CICC-ITOE)*, Jan 2010, pp. 50–53.

[11] Siwei Ma, Shiqi Wang, and Wen Gao, "Overview of ieee 1857 video coding standard," in *IEEE International Conference on Image Processing (ICIP)*, Sept 2013, pp. 1500–1504.

[12] Siwei Ma, Shiqi Wang, Qin Yu, Junjun Si, and Wen Gao, "Mode dependent coding tools for video coding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 990–1000, Dec 2013.

[13] Martin A. Fischler and Robert C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[14] G.J. Sullivan, J. Ohm, Woo-Jin Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," vol. 22, no. 12, pp. 1649–1668, Dec 2012.

[15] P. Helle, S. Oudin, B. Bross, D. Marpe, M.O. Bici, K. Ugur, J. Jung, G. Clare, and T. Wiegand, "Block merging for quadtree-based partitioning in HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1720–1731, Dec 2012.

[16] Liang Zhao, Xun Guo, Shawmin Lei, Siwei Ma, and Debin Zhao, "Simplified AMVP for high efficiency video coding," in *IEEE Visual Communications and Image Processing (VCIP)*, Nov 2012, pp. 1–4.

[17] Gianluca Francini, Skjalg Lepsøy, and Massimo Balestri, "Selection of local features for visual search," *Signal Processing: Image Communication*, vol. 28, no. 4, pp. 311–322, 2013.