

Research Article

Low-Complexity Saliency Detection Algorithm for Fast Perceptual Video Coding

Pengyu Liu and Kebin Jia

School of Electronic Information & Control Engineering, Beijing University of Technology, Beijing 100124, China

Correspondence should be addressed to Pengyu Liu; liupengyu.bjut@163.com

Received 27 September 2013; Accepted 5 November 2013

Academic Editors: M. Nappi and D. Tay

Copyright © 2013 P. Liu and K. Jia. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A low-complexity saliency detection algorithm for perceptual video coding is proposed; low-level encoding information is adopted as the characteristics of visual perception analysis. Firstly, this algorithm employs motion vector (MV) to extract temporal saliency region through fast MV noise filtering and translational MV checking procedure. Secondly, spatial saliency region is detected based on optimal prediction mode distributions in I-frame and P-frame. Then, it combines the spatiotemporal saliency detection results to define the video region of interest (VROI). The simulation results validate that the proposed algorithm can avoid a large amount of computation work in the visual perception characteristics analysis processing compared with other existing algorithms; it also has better performance in saliency detection for videos and can realize fast saliency detection. It can be used as a part of the video standard codec at medium-to-low bit-rates or combined with other algorithms in fast video coding.

1. Introduction

With the rapid developments of multimedia information processing and communication technology, video encoding has become the basic core technology of digital television, video conferencing, mobile media, 3D video coding, and so forth. During the past decades, in order to obtain video codec with low complexity, high quality, and high compression ratio, various technologies have been proposed for fast video coding [1].

Studies have shown that human visual system (HVS) is sensitive to the video scene perception and assigns different visual importance to different regions [2, 3]. Researchers have used advantages of visual attention in various multimedia processing applications such as image retargeting and video coding; the researching of saliency detection model for perceptual video coding is a hot topic. One of the key processing steps is to perform low-complexity calculations and obtain region of interest (ROI) in accordance with visual perception characteristics timely and effectively.

Up to now, saliency detection algorithms are widely used in extracting ROI in videos for various multimedia processing applications [4–8]. Moving zone detection technique for pixel precision is able to detect the moving foreground

area, but its complexity in calculation makes it not applicable in real-time encoding. Liu et al., [6] proposed the moving zone detection algorithm, but the algorithm mainly employs motion vector information, making it even not applicable in effective detection on moving object in global motion zone. Yuming et al., in [7] a video saliency detection algorithm based on feature contrast is proposed, but the computational efficiency needs to be further improved.

Those saliency detection algorithms mentioned previously are facing a common problem: they are not only time-consuming but computation-consuming as well. Those algorithms do not pay more consideration to the effect of the real-time coding performance because of the additional computation for visual perception analysis. Complex saliency detection algorithm will increase the computational burden of video encoder, which is not conducive to the video coding standards in real-time multimedia communication application.

In this paper, a fast saliency detection algorithm based on low-level encoding information and HVS is proposed. In order to simplify the visual perception analysis progress, this algorithm correlates the encoding information in video bit-stream with the visual perception characteristics. The spatial and temporal saliency detection is carried out by means



FIGURE 1: Tracking smooth movement targets (coastguard).

of MV. The prediction modes and other auxiliary coding information can save an amount of computing time in feature extraction for saliency detection. As almost no additional computation is increased to the video codec, the saliency detection computation complexity is lower compared with other existing algorithms. The saliency detection results are satisfied with the compared algorithms. So the proposed algorithm can reach the balance between the saliency detection accuracy and computational complexity.

This paper is organized as follows. Section 2 is an overview of the proposed framework and describes in detail each one of its subsystems. Section 3 gives the evaluation of the proposed algorithm, and in Section 4, some important conclusions are obtained and the further work is also introduced.

2. The Proposed Saliency Detection Algorithm

Motion is a highly salient feature which grabs one's attention and keeps it locked on important features and objects. Interest in motion perception has a long history and it can be considered as a relatively well established discipline [9]. Motion perception is one of the most important visual processing mechanisms. The visual information related to temporal motion would generate stronger response in HVS. HVS always pays more attention to the objects with smooth movement (as shown in Figure 1). Spatial contrast is the most basic visual processing mechanism in HVS; it is the prerequisite for HVS perception of spatial shape such as texture and object (as shown in Figure 2).

The response intensity of temporal motion visual information caused by HVS is larger compared with that of spatial motion visual information caused by HVS [10]. In the proposed algorithm, temporal saliency detection is performed firstly, and then spatial saliency detection is adopted in order to optimize the visual perception characteristics analysis results.

2.1. Temporal Saliency Analysis and Detection. As we know, moving objects always have larger MV in video frame. It can be found in Figure 3 that the coding regions with larger MV happen to be the ROI (such as head, face, shoulders, and arms); the coding regions with smaller MV or zero MV are always in the static background which could only arouse lower attention of HVS. To sum up, as a relatively high consistency exists between MV and visual attention, MV can be regarded as the temporal characteristics of visual perception. Many existing algorithms are used in order to get the motion feature for the motion saliency detection. These algorithms are only effective for videos with a static background.



FIGURE 2: Interest region of foreground goal (Suize).

In an ideal case, foreground moving object can generate nonzero MV. As there are no relative movements in the background region, it should produce zero MV in static background. However, in reality, nonzero MV noise would be generated randomly due to external change of illumination and internal change of video encoding parameters (such as change of quantization steps, motion search scheme, and rate distortion optimization algorithm). As a result, corresponding MV detection mechanism should be proposed to get rid of the interference of MV noise.

Besides the MV noise interfering in temporal saliency detection, the translation MV interfering from the background should be considered. As shown in Figure 4, in Foreman sequence, the foreground object is the human's head and shoulder. However, due to the rightward rotation of the camera, buildings on the right-hand side are moved into the scene, creating a globally distributed MV. In the same way, the horizontal displacement of camera causes obvious MV along the road side and the parked car in bus sequence. However, HVS is only interested in the moving bus on the road. Analogously, in Stefan sequence, the most interesting object is the tennis ball athlete which generated a lot of motion vector. But due to the moving of camera, there are amount of MV appeared on bleachers as well. Under these conditions with horizontal motion, the information of MV does not match the visual attention. Motion detection which is merely based on the size of MV can lead to errors in the temporal saliency judgment. As a result, corresponding MV detection mechanism needs to be formulated as well, in order to remove the interference caused by MV errors which are generated due to horizontal movement.

In order to improve the temporal detection accuracy, the MV noise filtering and the attenuating translation MV interference error should be added. At the same time, the complexity of processing procedure should be controlled strictly. Otherwise it will influence the real-time performance of the saliency detection algorithm.

2.1.1. Filtering MV Noise

(1) *Basic Principle of MV Noise Filtering.* Compression coding uses the correlation between adjacent macro blocks to



(a) Silent



(b) Salesman



(c) Paris

FIGURE 3: MV distribution in static background video frames.



(a) Foreman



(b) Bus



(c) Stefan

FIGURE 4: MV distribution in translation background video frames.

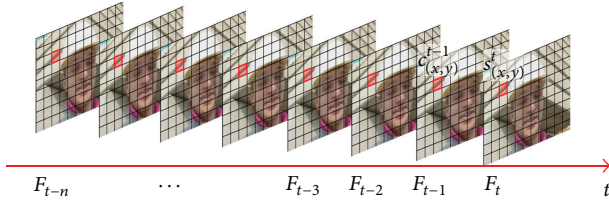


FIGURE 5: Schematic diagram of MBs with corresponding position in adjoining frames.

remove redundant information on spatial domain or temporal domain, in order to achieve a large amount of information with a small number of bits. In video coding framework based on blocks, the coded object is usually divided into several macro blocks or subblocks; then the macroblocks or subblocks that belong to an object should tend to have a similar motion vector or predictive coding mode and have similarity in structure [11]. Research shows that the moving objects in video sequence have the motion continuity and integrity. Motion continuity is reflected in there being a strong correlation between macro blocks in the current frame and the previous frame with corresponding position. Motion integrity reflects there existing great structural similarity between adjacent macro blocks in the same frame.

Therefore, based on the video sequence motion continuity and high temporal correlation, the MV of the macro block (MB) with the same position in the previous frame can provide very important prior information for the current MV of encoding MB.

In Figure 5, $s_{(x,y)}^t$ represents the current coded MB, t represents the current encoded video frame, and (x, y) is the position coordinates. \vec{V}_s is the MV of $s_{(x,y)}^t$. $c_{(x,y)}^{t-1}$ represents the MB in the previous frame with the same position coordinates as current coded MB and \vec{V}_c is the MV of $c_{(x,y)}^{t-1}$.

Through using large amounts of test sequences and statistics based on the H.264/AVC standard (JM18.7), it can be found that \vec{V}_s and \vec{V}_c have a high correlation.

Take Akiyo sequence and Foreman sequence as representations for gentle motion and active motion of the two kinds of video sequences; the quantization parameter (QP) is set as 28 and 32, respectively, through using full search prediction method, statistics the joint probability $p(\vec{V}_s | \vec{V}_c)$ of \vec{V}_s and \vec{V}_c . The statistical results are shown in Table 1.

From Table 1 statistics data, it can be found that if $\vec{V}_c = 0$, the probability of $\vec{V}_s = 0$ is more than 60%. If $\vec{V}_c \neq 0$, the probability of $\vec{V}_s \neq 0$ and belong to the \vec{V}_c ($1 \pm 10\%$) is nearly 80% for gentle motion sequence and more than 65% for active motion sequence. If $\vec{V}_c \neq 0$, the probability of $\vec{V}_s = 0$ is less than 20%.

The simulation results show that \vec{V}_c can be taken as an important basis for determining \vec{V}_s being MV noise or not. As there exists strong motion continuity and relativity of the moving object in a video sequence, in order to reduce the error judgment rate, in this paper, based on the average MV of reference region in previous frame, the basic principle of MV noise filtering is proposed as follows.

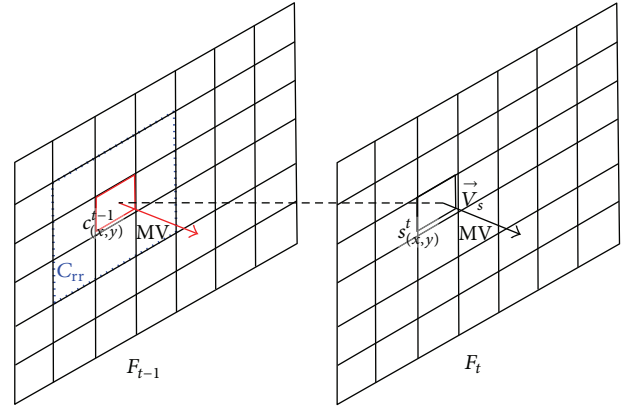


FIGURE 6: Schematic diagram of position relationship.

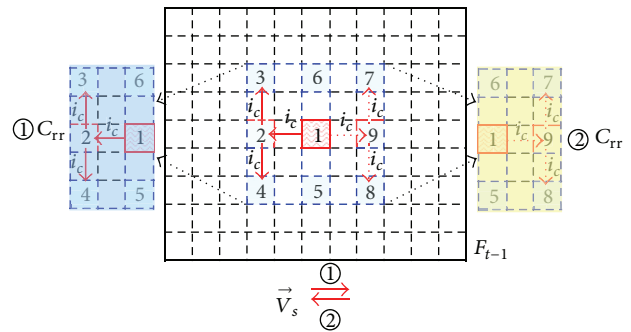


FIGURE 7: Schematic diagram of C_{rr} with horizontal motion.

If \vec{V}_s is generated in the current x encoding MB $s_{(x,y)}^t$, there is a high probability that a MV with similar direction and size exists in the reference region C_{rr} of the corresponding position in the previous frame. If there is no MV in C_{rr} , \vec{V}_s should be treated as MV noise and be filtered out. So how to determine the reference area C_{rr} is the key factor affecting the MV filtering results.

(2) Define Reference Region C_{rr} . As shown in Figure 6, $c_{(x,y)}^{t-1}$ is the MB which has the same position coordinates as MB $s_{(x,y)}^t$ in the previous frame. The rectangular area surrounded by dashed lines is defined as reference region C_{rr} .

According to the direction of \vec{V}_s (horizontal motion, vertical motion, and oblique motion), C_{rr} is determined as follows.

(i) C_{rr} with Horizontal Motion. As shown in Figure 7, if one takes horizontal motion towards the right direction of \vec{V}_s as an example, in the previous reference frame F_{t-1} , find MB-1, which has the same position coordinate with the current encoding MB signed $c(x, y)$.

Firstly, take MB-1 as the starting point, then perform horizontal motion of i_c macro blocks in the opposite direction of v_s , and get MB-2 signed $c(x, y - i_c)$.

Secondly, centered at MB-2, make vertical extension of i_c macro blocks both upwards and downwards to obtain two

TABLE 1: Probability of MV with the same position.

Video sequence	\vec{V}_c	QP = 28		\vec{V}_c	QP = 32		
		\vec{V}_s	$p(\vec{V}_s \vec{V}_c)$		\vec{V}_s	$p(\vec{V}_s \vec{V}_c)$	
Akiyo	0	0	66.38%	0	0	69.81%	
		$\neq 0$	33.62%		$\neq 0$	30.19%	
	$\neq 0$	0	11.92%	$\neq 0$	0	17.35%	
		$\in \vec{V}_c (1 \pm 10\%)$	76.42%		$\in \vec{V}_c (1 \pm 10\%)$	78.06%	
		Other values	11.66%			Other values	4.59%
Foreman	0	0	60.86%	0	0	64.72%	
		$\neq 0$	39.14%		$\neq 0$	35.28%	
	$\neq 0$	0	18.17%	$\neq 0$	0	19.69%	
		$\in \vec{V}_c (1 \pm 10\%)$	65.70%		$\in \vec{V}_c (1 \pm 10\%)$	68.02%	
		Other values	16.13%			Other values	12.29%

vertical vertices, that is, MB-3 signed $c(x - i_c, y - i_c)$ and MB-4 signed $c(x + i_c, y - i_c)$.

Last, determine the rectangular reference region ① and C_{rr} is designated by four macro blocks which are MB-3, MB-4, MB-5, and MB-6.

If \vec{V}_s is a horizontal motion towards the left direction, use the same method to determine the rectangular reference region ②; C_{rr} is designated and surrounded by four macro blocks which are MB-5, MB-6, MB-7, and MB-8.

In the above description, the position coordinates of MB-3 to MB-8 are given in Table 2.

In Table 2, i_c is defined as

$$i_c = \left\lceil \frac{|\vec{V}_{sx}|}{w_s} + 1 \right\rceil. \quad (1)$$

$|\vec{V}_{sx}|$ denotes the MV magnitude of the horizontal direction \vec{V}_s . w_s denotes the width of the current coding block. $\lceil \cdot \rceil$ represents the round numbers calculation.

(ii) C_{rr} with Vertical Motion. If \vec{V}_s make the vertical downward motion, the processing steps to determine reference region ③ C_{rr} are as follows.

Firstly, take MB-1 as the starting point, then perform vertical motion of j_c macro blocks in the opposite direction of \vec{V}_s and get MB-2 signed $c(x - j_c, y)$.

Secondly, centered at MB-2, make horizontal extension of j_c macro blocks both leftwards and rightwards to obtain two horizontal vertices: MB-3 signed $c(x - j_c, y - j_c)$ and MB-4 signed $c(x - j_c, y + j_c)$.

Last, determine the rectangular reference region ③; C_{rr} is designated and surrounded by four macro blocks which are MB-3, MB-4, MB-5, and MB-6 (as shown in Figure 8).

If \vec{V}_s make vertical upward motion, use the same method to determine the rectangular reference region ④; C_{rr} is designated and surrounded by four macro blocks which are MB-5, MB-6, MB-7, and MB-8.

In the above description, the position coordinates of MB-3 to MB-8 are given in Table 3.

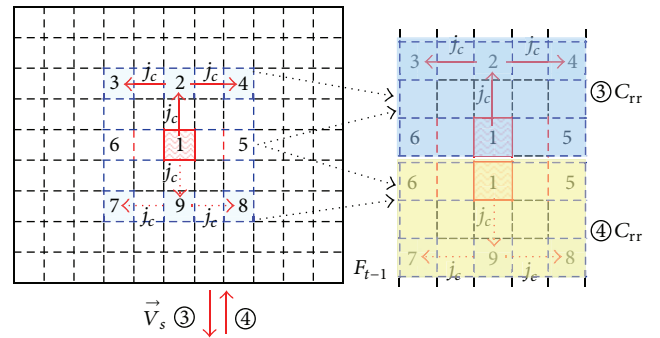

 FIGURE 8: Schematic diagram of C_{rr} with vertical motion.

TABLE 2: Position coordinates of vertex macroblocks with horizontal motion.

Macro blocks	Position coordinates
MB-3	$c(x - i_c, y - i_c)$
MB-4	$c(x + i_c, y - i_c)$
MB-5	$c(x + i_c, y)$
MB-6	$c(x - i_c, y)$
MB-7	$c(x - i_c, y + i_c)$
MB-8	$c(x + i_c, y + i_c)$

TABLE 3: Position coordinates of vertex macroblock with vertical motion.

Macro blocks	Position coordinates
MB-3	$c(x - j_c, y - j_c)$
MB-4	$c(x - j_c, y + j_c)$
MB-5	$c(x, y + j_c)$
MB-6	$c(x, y - j_c)$
MB-7	$c(x + j_c, y - j_c)$
MB-8	$c(x + j_c, y + j_c)$

In Table 3, j_c is defined as

$$j_c = \left\lceil \frac{|\vec{V}_{sy}|}{h_s} + 1 \right\rceil. \quad (2)$$

TABLE 4: Position coordinates of vertex macroblock with oblique motion.

Macro blocks	Position coordinates
MB-1	$c(x, y)$
MB-2	$c(x, y - i_c)$
MB-3	$c(x + j_c, y)$
MB-4	$c(x + j_c, y - i_c)$
MB-5	$c(x_c, y + i_c)$
MB-6	$c(x_c - j_c, y_c)$
MB-7	$c(x_c - j_c, y_c + i_c)$
MB-8	$c(x - j_c, y - i_c)$
MB-9	$c(x + j_c, y + i_c)$

$|\vec{V}_{sy}|$ denotes the MV magnitude of the vertical direction v_s . h_s denotes the height of the current coding block. $[\cdot]$ represent the round numbers calculation.

(iii) C_{rr} with Oblique Motion. If \vec{V}_s make the oblique motion, determine the reference regions of $\textcircled{3} C_{rr}$, $\textcircled{6} C_{rr}$, $\textcircled{7} C_{rr}$, and $\textcircled{8} C_{rr}$ in the same way according to the different motion directions of \vec{V}_s . C_{rr} is given in Figure 9.

The position coordinates of MB-1 to MB-9 are given in Table 4.

In Table 4, i_c and j_c are defined as formulas (1) and (2).

In the proposed algorithm, reference region C_{rr} is not stationary, and the area and position of C_{rr} are changed adaptively according to the size and direction of \vec{V}_s .

(3) The MV Noise Filtering Computational Formula. Consider

$$T_1(x, y, MV) = \begin{cases} 3, & \text{if } |\bar{V}_{rr}| = 0 \\ 2, & \text{else if } |\vec{V}_s| \geq |\bar{V}_{rr}| \\ T_2(x, y, MV) & \text{else.} \end{cases} \quad (3)$$

In formula (3), \bar{V}_{rr} is the averaged MV in C_{rr} ; it is defined as

$$\bar{V}_{rr} = \frac{\sum_{e \in C_{rr}} \vec{v}_{rr}}{\text{num}_{C_{rr}}}. \quad (4)$$

Here, \vec{v}_{rr} is the MV of MB in C_{rr} , $\text{num}_{C_{rr}}$ is the summation times. (x, y) denotes the position coordinates of current encoding MB.

If $|\bar{V}_{rr}| = 0$, consider \vec{V}_s is caused by MV noise and should be filtered out. \vec{V}_s is set as 0 and mark the current encoding block as 3, $T_1(x, y, MV) = 3$.

If $|\vec{V}_s| \geq |\bar{V}_{rr}|$, there exist obvious motion characteristics in the current encoding MB compared with the MBs in C_{rr} , and the current encoding block belongs to dynamic foreground region which should be marked as 2, $T_1(x, y, MV) = 2$.

Else, it means the current encoding MB has similar motion characteristics as its nearby MBs in C_{rr} , and the current encoding block's temporal saliency characteristics are undetermined. So the translational MV checking should be carried out further in order to distinguish whether current

encoding MB belongs to background region or foreground translation region.

2.1.2. Translational MV Checking. After MV noise filtering procedure, the translation MV interference attenuating step comes into consideration.

(1) Basic Principle of Translational MV Checking. In video coding based on the block matching, first step is to get the difference value between the best matching block and the original block, namely, the prediction error. If the prediction error value is smaller, after discrete cosine transform (DCT) transform the high frequency coefficients is less, then appearing probability of all zero quantized coefficients will be higher, and when number of coding bits is fewer, the higher the compression will be, which means the current block and the predicted block has more matching higher structural similarity, and the prediction coding effect is better.

In H.264/AVC standard, if one takes 4×4 subblock coding as example, the integer DCT can be described as

$$Y = (C_f X C_f^T) \otimes E$$

$$= \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & -2 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} [X] \begin{bmatrix} 1 & 2 & 1 & 1 \\ 1 & 1 & -1 & -2 \\ 1 & -1 & -1 & 2 \\ 1 & -2 & 1 & -1 \end{bmatrix} \right)$$

$$\otimes \begin{bmatrix} a^2 & \frac{ab}{2} & a^2 & \frac{ab}{2} \\ \frac{ab}{2} & \frac{b^2}{4} & \frac{ab}{2} & \frac{b^2}{4} \\ a^2 & \frac{ab}{2} & a^2 & \frac{ab}{2} \\ \frac{ab}{2} & \frac{b^2}{4} & \frac{ab}{2} & \frac{b^2}{4} \end{bmatrix}, \quad (5)$$

where $C_f X C_f^T$ is integer core transform, E is the constant scaling matrix, and \otimes represents matrix multiplication.

For a residual value of prediction error for 4×4 subblock $e(m, n)$, $0 \leq m, n \leq 3$, its transformation coefficient is $E(u, v)$ computational formula is as follows:

$$E(u, v)$$

$$= \sum_{m,n=0}^{3,3} e(m, n) \cdot \left[\frac{2.5C(u)}{\sqrt{2}} \times \cos \frac{(2m+1)u\pi}{8} \right]$$

$$\cdot \left[\frac{2.5C(v)}{\sqrt{2}} \times \cos \frac{(2n+1)v\pi}{8} \right]. \quad (6)$$

In formula (6), if $u, v = 0$, $C(u), C(v) = 1/\sqrt{2}$. If $u, v \neq 0$, $C(u), C(v) = 1$. $[\cdot]$ represents the rounding computation.

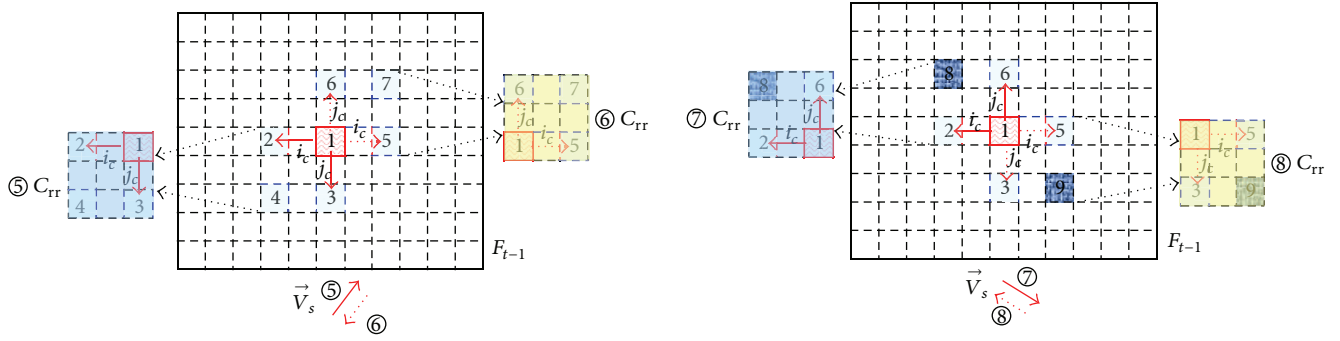


FIGURE 9: Schematic diagram of C_{rr} with oblique motion.

After getting $E(u, v)$, the computational formula of quantization coefficient $Z(u, v)$ is as follows:

$$|Z(u, v)| = (|E(u, v)| \times MF(u, v, QP) + f) \gg qbits$$

$$\text{Sign}(Z(u, v)) = \text{Sign}(E(u, v)) \quad (7)$$

$$qbits = 15 + \text{floor}\left(\frac{QP}{6}\right).$$

Here, $MF(u, v, QP)$ is the multiplier factor associated with QP . \gg stands for binary right shift operation. $f = 2^{qbits}/n$, $n = 3$ or 6 , corresponding intraframe coding block and interframe coding block, respectively.

If a quantized coefficient $Z(u, v)$ of an encoding block's $E(u, v)$ is equal to zero, it should be satisfied with the following conditions:

$$|E(u, v)| < \left\lfloor \frac{2^{qbits} - f}{MF(u, v, QP)} \right\rfloor. \quad (8)$$

The sum of absolute difference (SAD) of the 4×4 subblock can be obtained by adding prediction error absolute value of each region. Let

$$\text{SAD} = \sum_{m,n=0}^{3,3} |e(x, y)|. \quad (9)$$

So in the video coding standards based on the block-matching method, SAD is commonly used as the related function to measure the degree of correlation between the current encoding block and prediction block. The smaller value of SAD means there exiting stronger correlation between the two blocks, and they are more matchable. In this paper, the foreground translational region detection is on the basis of the of pixel region change detection theory [12].

(2) *Translational MV Checking Formula.* Let

$$T_2(x, y, MV) = \begin{cases} 1, & \text{if } \text{SAD}_{(x,y)} \geq \overline{\text{SAD}}_{S_c}, \\ 0, & \text{else.} \end{cases} \quad (10)$$

In formula (10), (x, y) represents the position coordinates of the encoding block. $\text{SAD}_{(x,y)}$ is the sum of absolute difference of the current encoding block and its corresponding

encoded block with the same position coordinates in previous frame. The value of $\text{SAD}_{(x,y)}$ can be described as variation degree of encoding blocks in two adjacent video frames. $\text{SAD}_{(x,y)}$ can be defined as follows:

$$\text{SAD}_{(x,y)} = \sum_{i=1}^M \sum_{j=1}^N |s(i, j) - c(i, j)|. \quad (11)$$

Here, $s(i, j)$ is the pixel value of the current encoding block. $c(i, j)$ is the pixel value of the corresponding block in previous frame. M, N denote the partition dimensions of current encoding block, respectively.

If the value of $\text{SAD}_{(x,y)}$ is high, it means that a great difference exists between the two adjacent frames. The current encoding block is considered in the foreground translational region under dynamic background condition and $T_2(x, y, MV)$ should be marked as 1.

If the value of $\text{SAD}_{(x,y)}$ is low, it means that a smaller difference exists between the two adjacent frames. And the current encoding block is considered in the background region and $T_2(x, y, MV)$ should be marked as 0.

(3) *Setting of Self-Adaptive Dynamic Threshold $\overline{\text{SAD}}_{S_c}$.* As there exists diversified motion degree in video sequences, different encoding parameters, especially quantization steps, can affect the code distortion and cause change in the value of $\text{SAD}_{(x,y)}$. How to measure $\text{SAD}_{(x,y)}$ value becomes one of the important factors affecting the performance of the proposed algorithm. Obviously using the fixed threshold will bring judgment error. In order to reduce the detection error caused by these uncertainties mentioned above, the proposed algorithm performs translational MV interference detection with a self-adaptive dynamic threshold $\overline{\text{SAD}}_{S_c}$, which can be determined by using the averaged $\text{SAD}_{(x,y)}$ value of all the encoding blocks in the background region in the previous frame. Let

$$\overline{\text{SAD}}_{S_c} = \frac{\sum_{(x,y) \in S_c} \text{SAD}_{(x,y)}}{\text{num}_{S_c}}. \quad (12)$$

Here, S_c represents the background region in the previous frame. $\sum_{(x,y) \in S_c} \text{SAD}_{(x,y)}$ is the summation of all the $\text{SAD}_{(x,y)}$ values for the encoding blocks enclosed in S_c . num_{S_c} is the summation times.

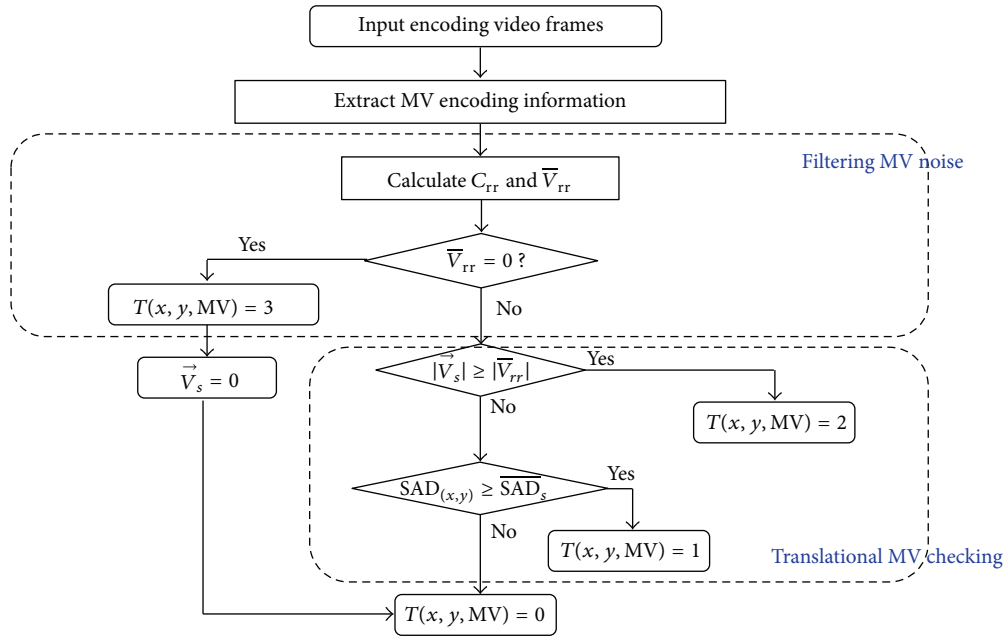


FIGURE 10: Flow chart of temporal saliency detection.

(4) *Temporal Saliency Flowchart*. In summary, the computational formula of temporal saliency analysis and detection is as follows:

$$T(x, y, MV) = \begin{cases} 3, & \text{if } |\bar{V}_{rr}| = 0 \\ 2, & \text{else if } |\vec{V}_s| \geq |\bar{V}_{rr}| \\ 1, & \text{else if } SAD_{(x,y)} \geq SAD_{S_c} \\ 0, & \text{else,} \end{cases} \quad (13)$$

where $T(x, y, MV) = 3$ means the current MV is MV noise; $T(x, y, MV) = 2$ means the current encoding block is belongs to the foreground dynamic region; $T(x, y, MV) = 1$ means the current encoding block is belongs to the foreground translational region; and $T(x, y, MV) = 0$ means the current encoding block is in the background region.

It should be pointed out that after filtering out the MV noise, $T(x, y, MV) = 3$, the current \vec{V}_s would be set to zero; the current encoding block belongs to background region, which should be marked with 0.

The flow chart of temporal saliency detection is shown in Figure 10.

According to the calculation procedure mentioned above, the current encoding frame can be sorted into temporal visual characteristic regions with different significance, based on the low-level encoding information \vec{V}_s of current encoding block and its motion vector relativity with adjacent blocks in C_{rr} (shown in Figure 11).

As the calculation of $SAD_{(x,y)}$ should be performed in interframe prediction mode decision and motion estimation, no additional calculation cost will be caused with the adoption of this method, so it is quite applicable in occasions with limited calculation resources.

2.2. Spatial Saliency Analysis and Detection. Because HVS is also sensitive to the change of spatial domain, in order to improve the visual perception of the analysis results, it needs to detect spatial saliency; analysis of the correlation between prediction mode and spatial visual features should also be performed. In the proposed algorithm, we take H.264/AVC standard as an example and discuss the correlation between prediction mode and visual spatial attention.

All the prediction modes of H.264/AVC coder are shown in Figure 12.

It has been verified in previous studies that the optimal prediction mode decision has the following rules [13].

In I-frame encoding of H.264/AVC standard, the smooth regions are suitable for using intra 16×16 , while regions with rich texture always select Intra 4×4 .

In P-frame encoding, the optimal prediction mode selection depends on the matching degree of encoded MB in forecasting, and the prediction mode selection results can describe the encoded MB's content richness and are consistent with the human visual selective attention.

As HVS is relatively nonsensitive to smooth background region, in H.264/AVC standard, the smooth regions usually choose the Intra 16×16 in I frame encoding or use macroblock prediction mode Inter 16 (skip, 16×16 , 16×8 , 8×16) in P frame encoding.

HVS usually assigns higher visual importance to figures in foreground regions with abundant texture features and moving objects; therefore, those regions mentioned always select Intra 4×4 in I frame encoding or use subblock prediction mode Inter 8 (8×8 , 8×4 , 4×8 , 4×4) in P frame encoding (shown in Figure 13).

Although it is small probability event that using Intra mode in P-frame coding, once Intra mode is selected, means

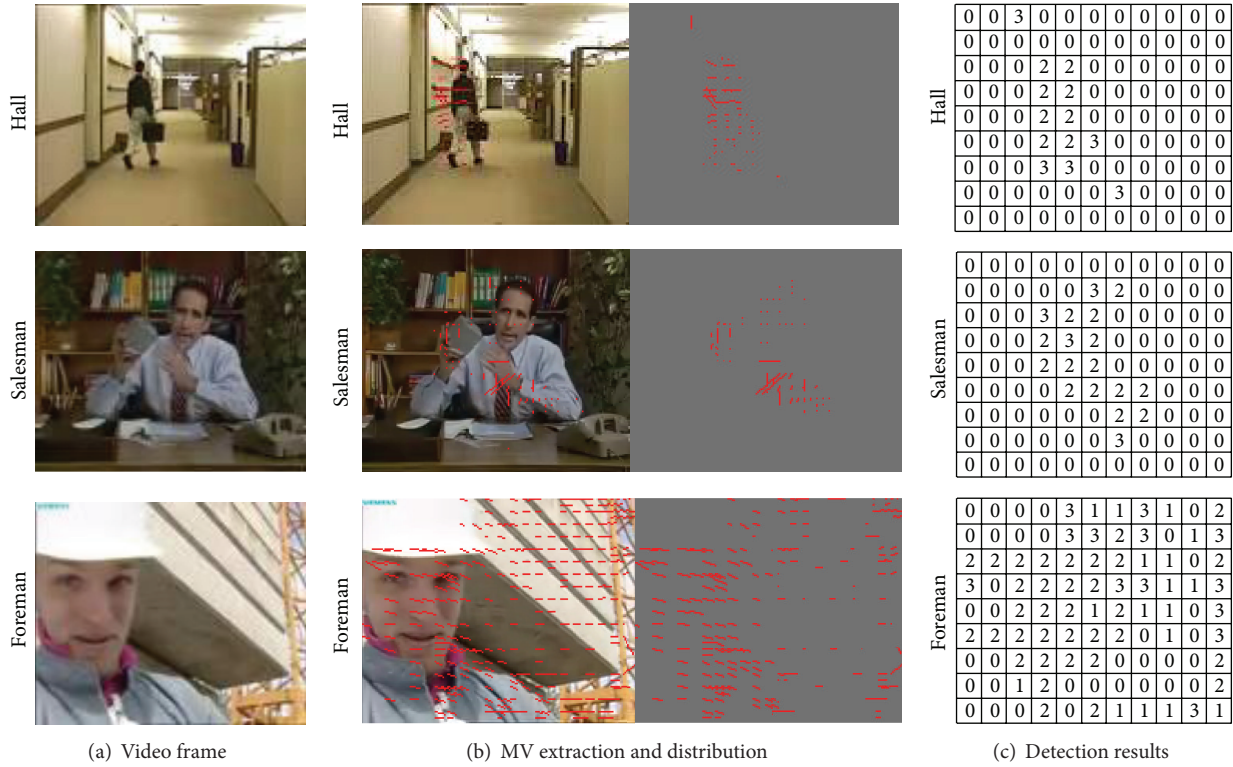


FIGURE 11: Schematic diagram of temporal saliency detection results based on MV.

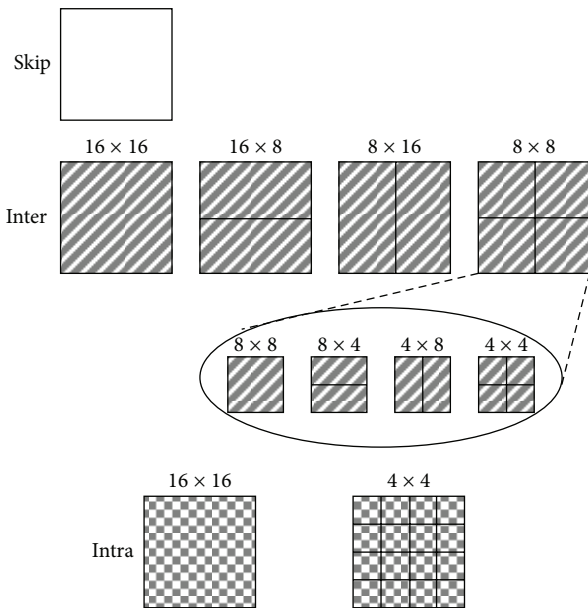


FIGURE 12: Prediction modes in H.264/AVC.

there appears new information or the encoding content is varied greatly in current frame compared with previous frame. It can be found that, in Figure 13(d), there are 6 MBs (with red dots mark) which select Intra 4 × 4 as the optimal prediction mode; because the woman raised right arm

suddenly, the moving arm is just the ROI with higher human visual attention.

There is high consistency between prediction mode decision results and visual attention. Therefore, in the proposed algorithm, the prediction mode is regarded as spatial characteristic of visual perception analysis.

According to the analysis above, current encoding frame can be sorted into spatial visual characteristic regions with different significance according to optimal prediction mode decision results. The spatial saliency detection computational formula is as follows:

$$S(x, y, Mode) = \begin{cases} 2, & Mode_p \in \{Intra\} \\ 1, & Mode_p \in \{Inter\ 8\} \text{ or } Mode_I \in Intra\ 4 \times 4 \\ 0, & Mode_p \in \{Inter\ 16\} \text{ or } Mode_I \in Intra\ 16 \times 16. \end{cases} \quad (14)$$

Mode_p, Mode_I represent the optimal prediction modes selected by the current MB in P-frame and I-frame coding, respectively.

If Mode_p ∈ {Intra}, means in P-frame coding, then Intra 4 × 4 or Intra 16 × 16 is selected as the optimal prediction mode, the spatial saliency is high, the current encoding block belongs to the human visual sensitive region, and it can be expressed as S(x, y, Mode) = 2.

If Mode_p ∈ {Inter 8} or Mode_I ∈ Intra 4×4, which means the current encoding block takes the subblock prediction mode (8 × 8, 8 × 4, 4 × 8, 4 × 4) as the optimal prediction mode

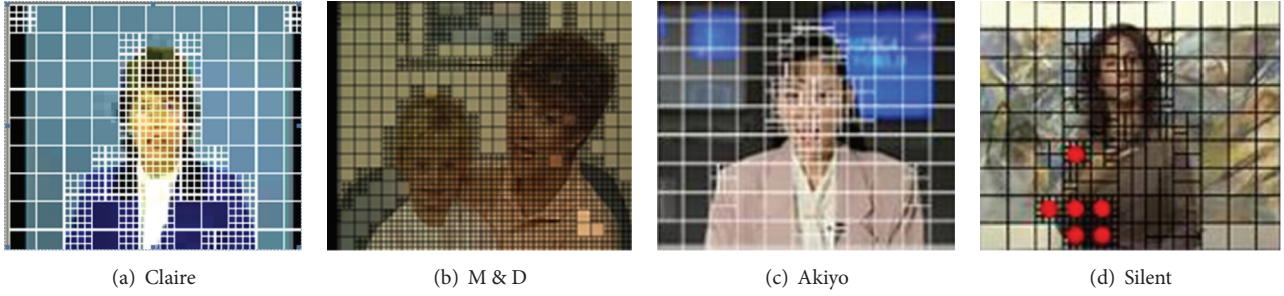


FIGURE 13: Relationship between prediction modes with visual attention. ((a), (b)) I frame encoding, ((c), (d)) P frame encoding.

in P-frame coding or uses Intra 4×4 in I-frame coding, then the current encoding block has abundance of texture feature or changes in spatial domain, the current block belongs to attention region, the spatial saliency is high, and then it can be expressed as $S(x, y, \text{Mode}) = 1$.

If $\text{Mode}_p \in \{\text{Inter } 16\}$ or $\text{Mode}_I \in \text{Intra } 16 \times 16$, which means the current encoding block is smooth and belong to the visual nonsensitive region, then the spatial changes are slight, the current block has low spatial visual characteristics significance, it belongs to nonsignificant region, and it can be signed with $S(x, y, \text{Mode}) = 0$.

2.3. Combination of the Spatiotemporal Saliency Detection. In this paper, according to the spatiotemporal saliency detection results, we define video region of interest (VROI). Let

$$\text{VROI}(x, y, \text{MV}, \text{Mode}) = T(x, y, \text{MV}) \parallel S(x, y, \text{Mode}). \quad (15)$$

The computational formula can be expressed as

$$\text{VROI}(x, y, \text{MV}, \text{Mode}) = \begin{cases} 5, & S(x, y, \text{Mode}) = 2 \\ 4, & T(x, y, \text{MV}) = 2 \parallel S(x, y, \text{Mode}) = 1 \\ 3, & T(x, y, \text{MV}) = 1 \parallel S(x, y, \text{Mode}) = 1 \\ 2, & (T(x, y, \text{MV}) = 2 \text{ or } T(x, y, \text{MV}) = 1) \parallel \\ & S(x, y, \text{Mode}) = 0 \\ 1, & (T(x, y, \text{MV}) = 0 \text{ or } T(x, y, \text{MV}) = 3) \parallel \\ & S(x, y, \text{Mode}) = 1 \\ 0, & (T(x, y, \text{MV}) = 0 \text{ or } T(x, y, \text{MV}) = 3) \parallel \\ & S(x, y, \text{Mode}) = 0. \end{cases} \quad (16)$$

In this algorithm, according to the calculation number of $T(x, y, \text{MV})$ and $S(x, y, \text{Mode})$, the priority level of VROI is divided into 6 grades, from high to low being 5~0. For example, if the current MB uses Intra mode in P-frame coding ($S(x, y, \text{Mode}) = 2$), this means that the MB is in the visual sensitive region and has the highest visual attention; it can be signed with $\text{VROI}(x, y, \text{MV}, \text{Mode}) = 5$, and so on.

The proposed algorithm framework is as depicted in Figure 14.

3. The Algorithm Performance Evaluation

3.1. Test Platform. In this paper, three existing algorithms are used to do the comparison experiment [4, 6, 7]. The experimental environment is set as Table 5.

We use 10 typical test sequences in multiple formats which include various types (such as 176×144 , 352×288 , and 416×240) of video with different scenes, motion, and flatness, separately, such as videos in daytime and nighttime, sports videos, news television, broadcast, and video surveillance.

3.2. Saliency Detection Results. According to the visual perception characteristics analysis and the saliency detection procedure mentioned previously, the VROI marking results are shown in Figure 15.

In the output saliency detection results, the MB luminance values are proportional to the priority level of $\text{VROI}(x, y, \text{MV}, \text{Mode})$. The region with higher visual sensitive, the corresponding MB luminance value is higher. In Figure 15, MVD is the motion vector diagram, and PMD is the prediction mode diagram. The detection results have good consistency with human visual system.

3.3. Algorithm Complexity Analysis. The computation time and the similar measure method are adopted to evaluate the performance of the algorithm [13]. In the similar measure method, Kullback-Leibler (KL) distance is used to measure the similarity between the saliency distributions at human saccade locations and random locations as

$$\text{KL}(H, R) = \frac{\sum_k h_k \log(h_k/r_k) + \sum_k r_k \log(r_k/h_k)}{2}. \quad (17)$$

Here H and R are saliency distributions at human saccade locations and random locations with probability density functions h_k and r_k , respectively.

The saliency detection algorithm with the higher KL distance can discriminate human saccade locations from the random locations more easily, and this means better performance in saliency detection for videos [14].

Statistical data in Table 6 show that the proposed algorithm can enhance the timeliness of calculation and the performance in video saliency detection markedly. Compared with [4, 6, 7], the calculation time for VROI detection can be saved up to 10.69%, 14.66%, and 5.29%; at the same time, the KL distance is increased by 0.50, 1.02, and 0.24, respectively.

TABLE 5: Experimental environment.

Computer hardware	P4 @ 1.6 GHz and 2 G RAM	
Experimental software platform	H.264/AVC (JM18.7), Visual C++, Windows 2003	
	Experimental parameters	
Encoding frames	Frame rate	Gop
100	30 f/s	IPPP
Entropy encoding type	QP	Search range
CAVLC	32	± 16 pixels
Reference frames number	Hadamard transform	RDO
5	On	On

TABLE 6: Algorithms performance comparison results.

Sequences	Comparison algorithms	ΔC time (%)	ΔKLD
Foreman	[4]	-11.14	+0.59
	[6]	-12.88	+1.02
	[7]	-6.91	+0.32
Hall	[4]	-10.22	+0.31
	[6]	-13.75	+0.68
	[7]	-6.55	+0.21
Salesman	[4]	-12.25	+0.47
	[6]	-17.24	+0.96
	[7]	-7.12	+0.36
Paris	[4]	-12.26	+0.71
	[6]	-18.77	+1.28
	[7]	-4.90	+0.16
Silent	[4]	-7.34	+0.33
	[6]	-12.45	+0.60
	[7]	-2.76	+0.19
Akiyo	[4]	-8.19	+0.94
	[6]	-10.76	+1.56
	[7]	-4.81	+0.37
Bus	[4]	-9.75	+0.38
	[6]	-20.01	+1.49
	[7]	-5.47	+0.27
Mobisode2	[4]	-14.21	+0.31
	[6]	-10.78	+0.47
	[7]	-4.07	+0.05
Stefan	[4]	-13.38	+0.72
	[6]	-19.65	+1.34
	[7]	-5.78	+0.36
Container	[4]	-8.11	+0.22
	[6]	-10.27	+0.78
	[7]	-4.56	+0.11
Average	[4]	Δ Computing time (%)	Δ KL distance
		-10.69	+0.50
	[6]	Δ Computing time (%)	Δ KL distance
	-14.66	+1.02	
	[7]	Δ Computing time (%)	Δ KL distance
		-5.29	+0.24

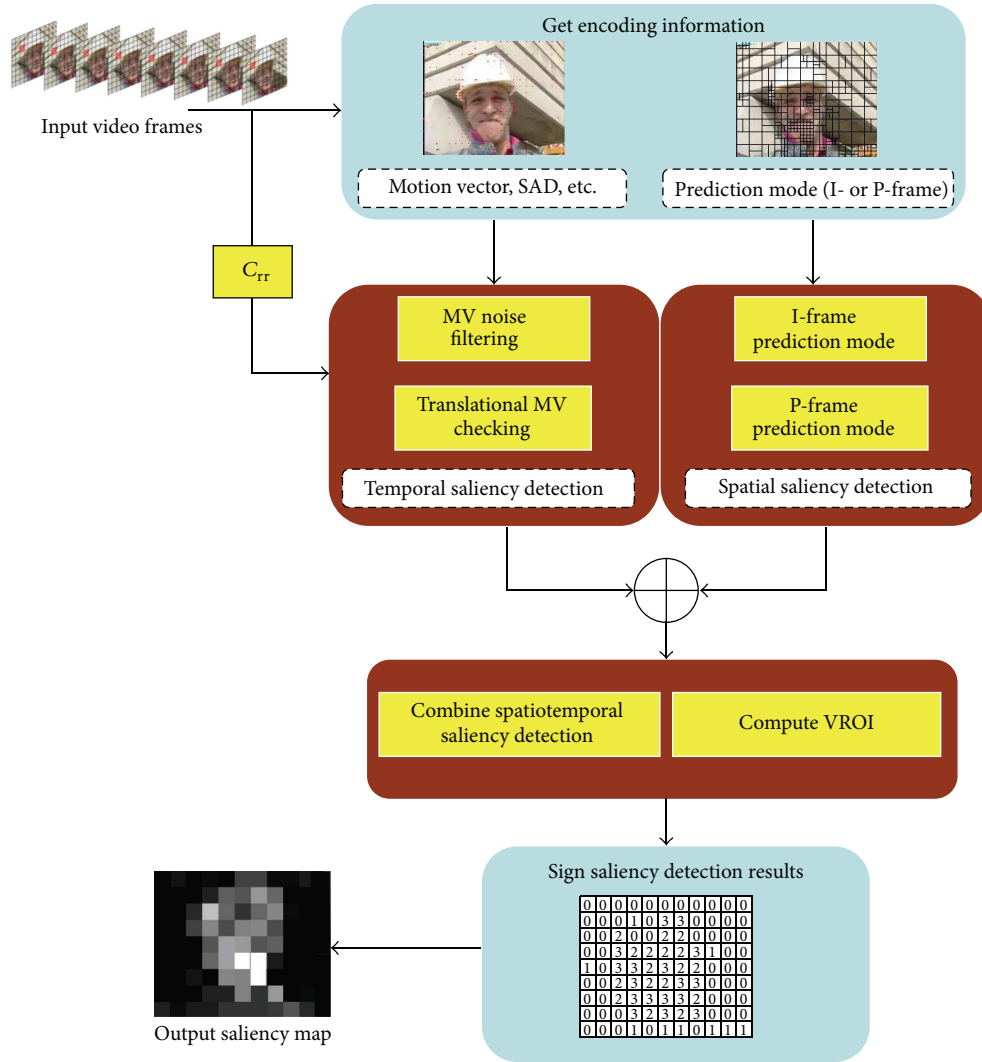


FIGURE 14: Diagram of proposed saliency detection algorithm framework.

Especially for Bus, Stefan, Foreman, and Paris sequences, which contain a large number of global motion vectors or rich texture of background, the proposed algorithm can analyze visual perception characteristics and extraction VROI fast and accurately.

In Table 6, mathematical symbol “-” denotes decrease and “+” denotes increase. ΔC Time (%), ΔKLD are defined as follows formula in (18) and (19), respectively:

$$\Delta C \text{ Time (\%)} = \frac{\text{ComputingTime}_{\text{compared}} - \text{ComputingTime}_{\text{proposed}}}{\text{ComputingTime}_{\text{compared}}} \times 100\%, \tag{18}$$

$$\Delta KLD = KLDistance_{\text{proposed}} - KLDistance_{\text{compared}}. \tag{19}$$

In Table 6, the average computational time is the shortest, and the average KL distance of the proposed algorithm is the

largest. This means the proposed algorithm can control the computational complexity strictly and discriminate human saccade locations from random locations more quickly and accurately than the other ones. The experiment results demonstrate that the performance of the proposed algorithm is the best among these compared ones in video saliency detection.

4. Conclusion

In this paper, the interdependency between video encoding information and HVS characteristics is studied; it proposes a video saliency detection algorithm based on visual perception characteristics analysis and low-layer encoding information which can get from the bit-stream directly. The simulation results show that the proposed algorithm has better performance than other existing ones. It can filter out the motion vector noise, weaken the interference of translational motion vector and get rid of visual redundancy, and it can be used in the detection of visual perception


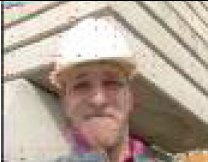
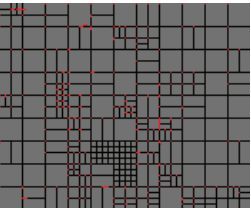
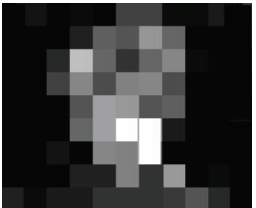



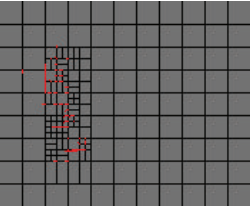

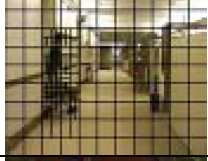


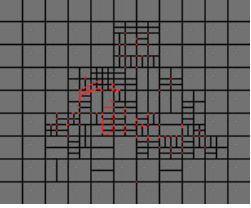

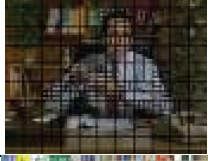


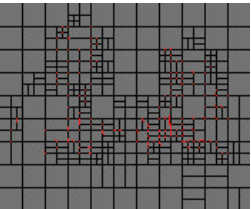


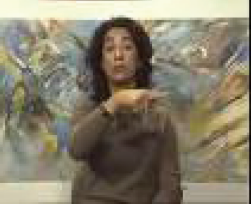

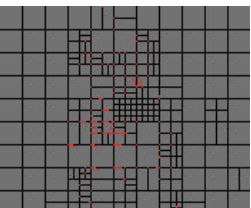
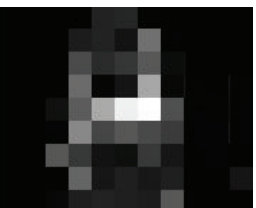

Video frame	Encoding information		Saliency detection results	
 <p>Foreman_qcif (17th)</p>	MVD			
	PMD			
 <p>Hall_qcif (13th)</p>	MVD			
	PMD			
 <p>Salesman_qcif (7th)</p>	MVD			
	PMD			
 <p>Paris_qcif (45th)</p>	MVD			
	PMD			
 <p>Silent_qcif (39th)</p>	MVD			
	PMD			

FIGURE 15: Continued.

Video frame	Encoding information		Saliency detection results	
 <p>Akiyo_cif (68th)</p>	MVD			
	PMD			
 <p>Bus_cif (42th)</p>	MVD			
	PMD			
 <p>Mobisode2.416 × 240 (86th)</p>	MVD			
	PMD			
 <p>Stefan_cif (6th)</p>	MVD			
	PMD			
 <p>Container_cif (30th)</p>	MVD			
	PMD			

FIGURE 15: The saliency detection results of the proposed algorithm.

characteristics analysis and saliency detection fast and effectively. The complication of the proposed algorithm is low, and its detection results are more consistent with HVS compared with other existing algorithms. It can be used conveniently in many Internet-based multimedia applications such as video retrieval based on ROI and video quality assessment. It can also be applied to video coding standards, such as HEVC and H.264/AVC.

In the future, the various multimedia applications of the proposed video saliency detection algorithm combined with fast video coding technologies can realize fast video coding based on HVS for the latest video coding standard HEVC, and saliency detection technique can be taken as part of the video standard codec at medium-to-low bit-rates.

Acknowledgment

The research work is supported by the National Key Technology R&D Program of China with Grant no. 2011BAC12B03, the National Natural Young Science Foundation of China with Grant no. 61100131, and Beijing City Board of Education Project with Grant no. KM201110005007.

References

- [1] K. Andreas, G. Alexander, T. Michael, E. Marko, and S. Thomas, "Adaptive global motion temporal filtering for high efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1802–1812, 2012.
- [2] M. Jiang, Z. Ma, X. Niu, Y. Yang, and Y. Yang, "An HVS-based JPEG image watermarking method for copyright protection," *International Journal of Digital Content Technology and its Applications*, vol. 5, no. 10, pp. 11–19, 2011.
- [3] Y. Wei, C. Lap-Pui, and R. Susanto, "Joint rate allocation for statistical multiplexing in video broadcast applications," *IEEE Transactions on Broadcasting*, vol. 58, no. 3, pp. 417–427, 2012.
- [4] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.
- [5] Y. Liu, Z. G. Li, Y. C. Soh, and M. H. Loke, "Conversational video communication of H.264/AVC with region-of-interest concern," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '06)*, pp. 3129–3132, October 2006.
- [6] Z. Liu, Z. Zhang, and L. Shen, "Moving object segmentation in the H.264 compressed domain," *Optical Engineering*, vol. 46, no. 1, Article ID 017003, 2007.
- [7] F. Yuming, L. Weisi, W. Lin, C. Zhenzhong, T. Chia-Ming, and L. Chia-Wen, "Video saliency detection in the compressed domain," in *Proceedings of the 20th ACM International Conference on Multimedia*, pp. 697–700, Nara, Japan, 2012.
- [8] I. Nevrez, L. Weisi, and F. Yuming, "A saliency detection model using low-level features based on wavelet transform," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 96–105, 2013.
- [9] J. Culham, S. He, S. Dukelow, and F. A. J. Verstraten, "Visual motion and the human brain: What has neuroimaging told us?" *Acta Psychologica*, vol. 107, no. 1-3, pp. 69–94, 2001.
- [10] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 631–637, June 2005.
- [11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [12] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 294–307, 2005.
- [13] P. Liu, K. Jia, and Y. Zhang, "A layered structure prediction method for mode decision in video encoding," in *Proceedings of the 8th International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 407–410, Piraeus-Athens, Greece, July 2012.
- [14] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Proceedings of the Neural Information Processing Systems (NIPS '06)*, 2006.

