IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 5, 2017, accepted May 22, 2017, date of publication June 8, 2017, date of current version June 27, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2712789

Survey of Spatio-Temporal Interest Point Detection Algorithms in Video

YANSHAN LI[1], RONGJIE XIA[1], QINGHUA HUANG[2,3,4], WEIXIN XIE[1], AND XUELONG LI[5]

[1]ATR National Key Laboratory of Defense Technology, Shenzhen University, Shenzhen 518060, China
[2]College of Information Engineering, Shenzhen University, Shenzhen 518060, China
[3]School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China
[4]Center for Optical Imagery Analysis and Learning (OPTIMAL), School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China
[5]State Key Laboratory of Transient Optics and Photonics, Center for Optical Imagery Analysis and Learning (OPTIMAL), Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China

Corresponding author: Qinghua Huang (qhhuang@scut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61401286, Grant 61372007, and Grant 61571193, in part by the Foundation for Distinguished Young Talents in Higher Education of Guangdong under Grant 2014KQNCX132, in part by the Shenzhen Science and Technology Project under Grant JCYJ20160307143441261, in part by the National Defense Preliminary Research Project under Grant 9140C80050215 0C80341, in part by the Guangzhou Key Lab of Body Data Science under Grant 201605030011, and in part by the Guangdong Provincial Science and Technology Program International Collaborative Projects under Grant 2014A050503020).

ABSTRACT Recently, increasing attention has been paid to the detection of spatio-temporal interest points (STIPs), which has become a key technique and research focus in the field of computer vision. Its applications include human action recognition, video surveillance, video summarization, and content-based video retrieval. Amount of work has been done by many researchers in STIP detection. This paper presents a comprehensive review on STIP detection algorithms. We first propose the detailed introductions and analysis of the existing STIP detection algorithms. STIP detection algorithms are robust in detecting interest points for video in the spatio-temporal domain. Next, we summarize the existing challenges in the STIP detection for video, such as low time efficiency, poor robustness with respect to camera movement, illumination change, perspective occlusion, and background clutter. This paper also presents the application situations of STIP and discusses the potential development trends of STIP detection.

INDEX TERMS Video, spatio-temporal interest point (STIP), local invariant feature, STIP detection algorithm.

I. INTRODUCTION

Spatio-temporal interest point (STIP) is one type of local invariant feature for video. Local invariant feature can resist changes such as rotation, scale variations, affine transform and view point change. The most frequently used features include Harris corner [1], Scale Invariant Feature Transform (SIFT) [2], [3] and Speeded Up Robust Features (SURF) [4]. They have been widely and successfully applied to object detection and recognition, image registration, image classification and image analysis [5]–[11]. Based on the local invariant feature of gray or color image, STIP was proposed by Laptev *et al.* in 2003 firstly [12]. STIPs are points with significant local variation of the pixel intensity in the spatio-temporal domain of the video volume. Compared with global features, STIP is robust and able to resist different variations of video, such as geometric transformation, perspective transformation, illumination variation and convolution transformation. Moreover, STIP can be detected directly from video to describe moving objects, without the need for background modeling and foreground segmentation. In recent years, STIP has become one of the research hotspots in the area of video analysis due to the advantages mentioned above. Many excellent algorithms have been proposed [12]–[16] and widely used in various applications, including human action recognition [17]–[20], video surveillance [21]–[23], video summarization [24]–[26] and content-based video retrieval [27]–[29].

II. STIP DETECTION ALGORITHMS IN VIDEO

Corner and blob are two categories of local feature in 2D image. Corner is the intersection point of two edges, and blob refers to the region with brightness difference or color difference compared to the surrounding regions. A blob is the region of an image in which some properties are constant or approximately constant. Through comprehensive comparison of the state-of-the-art research achievements, we also divide STIP into spatio-temporal corner (STC for short) and spatio-temporal blob (STB for short) two categories.

A. STC AND ITS DETECTION ALGORITHMS

Defining spatio-temporal corners are image regions with significant variation of the local gradient in orthogonal directions, i.e. x, y and t. In practical, most of the so-called corner detection algorithms not merely detect corner points, but detect points which have specific coordinates and some mathematical characteristics (*e.g.,* the local maximum or minimum of the gray value) and some gradient features. STC detection algorithms detect corners by calculating the curvature and gradient of points, which have the advantage of good robustness. Their key procedures are as follows. Firstly, a strength response function is usually given to calculate the intensity of each point. Then, candidate feature points in the spatio-temporal volume are obtained by non-maximum suppression. Finally, the real STC points are obtained by imposing some constraints.

Harris3D detector firstly constructs the linear scale-space representation L by convolving the input signal f with an anisotropic Gaussian kernel g related independent spatial variance σ^2 and temporal variance τ^2 [12]. And g is shown in Eq.(1).

$$\mathrm{g}(x, y, t; \sigma^2, \tau^2) = \frac{1}{\sqrt{(2\pi)^3\sigma^4\tau^2}} e^{-\frac{x^2+y^2}{2\sigma^2} - \frac{t^2}{2\tau^2}} \tag{1}$$

Then it uses g and the spatio-temporal matrix which is composed by the second-order derivatives of L to compute a spatio-temporal second-order matrix μ at each point of the video as Eq.(2).

$$\mu = \mathrm{g}(x, y, t; \sigma^2, \tau^2) * \begin{bmatrix} L_{\mathrm{xx}} & L_{\mathrm{xy}} & L_{\mathrm{xt}} \\ L_{\mathrm{yx}} & L_{\mathrm{yy}} & L_{\mathrm{yt}} \\ L_{\mathrm{tx}} & L_{\mathrm{ty}} & L_{\mathrm{tt}} \end{bmatrix} \tag{2}$$

where, L_{xx}, L_{xy}, L_{xt}, L_{yx}, L_{yy}, L_{yt}, L_{tx}, L_{ty}, L_{tt} are the second-order derivatives of L.

Then, spatio-temporal Harris corner response function H is designed as Eq.(3). STIPs of f can be found by searching local positive spatio-temporal maximum of H.

$$H = \det(\mu) - k \cdot trace^3(\mu) \tag{3}$$

where det (μ) is the determinant of μ, and $trace(\mu)$ is the trace of μ.

Harris3D spatio-temporal corner detector is extended from Harris corner detector [1] of gray image and it detects local structure where the image values have significant variations in both the spatial domain and the temporal domain. The detector is efficient to extract STIPs and it can detect STIPs in scenes with occlusions and dynamic cluttered backgrounds. It has scale adaptability [30] due to the Gaussian filter, and it uses local maximization of the normalized spatio-temporal Laplacian operator to realize the scale selection in the spatio-temporal domain. However, since Gaussian filter is non-causal but the variation of the video signal in the temporal domain is causal [15], using Gaussian filter in the temporal domain is not appropriate in the respect of using the causality of video data effectively. Moreover, Harris3D detector separately and repeatedly detects each candidate STIP, making it a time-consuming iterative procedure. Because the iterative procedure often diverges, it detects only a sparse set of features for keeping the computation time under the control.

In order to solve the problem that the non-causal Gaussian filter does not apply to causal time domain of video, Dollár *et al.* [13] proposed a Cuboid detector whose response function R is defined as Eq.(4).

$$R = (f * g * h_{ev})^2 + (f * g * h_{od})^2 \tag{4}$$

where, g is a two-dimensional Gaussian kernel applied in the spatial domain of video, as shown in Eq.(5).

$$g(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp(-(x^2 + y^2)/2\sigma^2) \tag{5}$$

The h_{ev} and h_{od} are a pair of orthogonal one-dimensional Gabor filter applied to the time domain of video, and they are defined as follows.

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2} \tag{6}$$

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2} \tag{7}$$

STIPs are finally obtained by calculating the value of R of each point and using the non-maximum suppression. After detecting STIPs, they expand each STIP into a cuboid, and finally achieve cuboid prototypes by k-means clustering. It can detect abundant STIPs because 2D Gaussian filters are applied spatially and 1D Gabor filters are applied temporally. Gabor filter, a linear separable filter, is sensitive to the periodic change of pixel intensity of video, hence the method can detect object with significant motion, especially periodic motion. However, the intensity of R based on Gabor filter is linear with the boundary distinction and motion complexity, making STIPs unable to reflect the motion change, such as translational motion. Its another shortage is that it does not take the advantage of scale invariance and hence the scale parameters of the R must be set manually before detecting STIPs, making it unable to track the change of image scale. Therefore, the detector only works effectively in the case of a fixed scale.

In order to improve the detector to be scale-invariant, Ke *et al.* [31] proposed the Volumetric STIP detector, Oikonomopoulos *et al.* [32] presented a detector by utilizing the entropy information of optical flow, and Laptev *et al.* [33] used the local motion of objects to detect STIPs.

The algorithms mentioned above have good performance with respect to feature detection. However, most of them use Gaussian filters in both the spatial domain and the temporal domain, which ignores the uncertainty in temporal correlation. It not only causes the spatio-temporal structure to be confused and blurred, but also leads to low accuracy and sensitivity when those algorithms are detecting motions in video. Especially, when the scene presents dynamic backgrounds or unknown ego disturbances (*e.g.,* camera shaking or movement, perspective transformation), these drawbacks are more apparently exposed. Therefore, Shabani *et al.* [15] proposed to generate scale space for detecting STIPs in real-world video with non-linear scale-space filtering, as shown in the following equation.

$$\partial I(x, y, t_i, s)/\partial s = div(g_{sp}(\left|\nabla I_{sp}\right|) \cdot \nabla I) + g_{tp}(\left|\nabla I_{tp}\right|) \cdot \nabla I \tag{8}$$

where, $I(x, y, t_i, s)$ is the smoothed image of the video frame $I(x, y, t_i)$ with the scale parameter s, the edge-stopping function g_{sp} and salient-motion-stopping function g_{tp} are non-linear functions which steer the energy redistribution in the scale-space filtering. g_{sp} and g_{tp} are defined as follows.

$$\mathrm{g}_{sp}(\left|\nabla I_{sp}\right|) = 1/(1 + (\left|\nabla I_{sp}\right|/\lambda_{sp})^2) \tag{9}$$

$$\mathrm{g}_{tp}(\left|\nabla I_{tp}\right|) = 1/(1 + (\left|\nabla I_{tp}\right|/\lambda_{tp})^2) \tag{10}$$

where, λ_{sp} and λ_{tp} are the threshold factors controlling the important level of the gradient in the spatio-temporal domain. The method is implemented as follows. Firstly, the non-linear spatio-temporal scale-space filtering is used for smoothing each frame of video. Then, the difference between the current frame and the prior frame is computed to obtain the non-linear counterpart of the Laplace of Gaussian (LoG). Finally, the detected STIPs are those points with local maximum in the spatio-temporal volume of non-linear LoG. Because the incorporation of the spatial and temporal feedback in the stopping function (g_{sp} and g_{tp}), the scale-space filtering is non-linear. Besides, this method puts forwards a new model to handle the uncertainty of temporal correlation due to the unknown camera movement. It not only prevents disturbances produced by noises and cluttered backgrounds, but also removes ego disturbances caused by unknown camera movement. However, blurring and dislocation usually appear at the edges of moving objects. Furthermore, this method is merely applied to the video data with slight camera shaking or moving. Therefore, it detects a small amount of STIPs on moving objects in the case of video with highly cluttered backgrounds or the severe camera movement. Therefore, Liu *et al.* [34] improved the non-linear anisotropic diffusion filter [35] to detect STIPs in scenes with cluttered backgrounds caused by the camera movement. Non-linear anisotropic diffusion filters smooth noise and textureless regions in cluttered backgrounds efficiently and extract the edges of the object. Therefore, this algorithm can suppress disturbances of cluttered backgrounds and the camera movement. The detected STIPs are those points located in the moving objects and are more robust. However, it is time-consuming because the non-linear anisotropic diffusion filters diffuse at varied degrees.

For the same purpose, i.e., to solve a series of problems in complex scenes, Chakraborty *et al.* [16] presented a novel approach called selective STIP detector by imposing local and temporal constraints. Firstly, they utilized Harris corner detector [1] to detect the first set of spatial interest points, many of which belong to backgrounds. Then, they defined a surround suppression mask (SSM) $t_\sigma(x, y)$ of each interest point $C_\sigma(x, y)$ as the weighted sum of gradient weighting factors $\Delta_{\Theta,\sigma}(x, y, x - u, y - v)$ as follows,

$$\begin{aligned} \mathrm{t}_\sigma(x, y) = \iint_\Omega & C_\sigma(x - u, y - v) \\ & \times \Delta_{\Theta,\sigma}(x, y, x - u, y - v)dudv \end{aligned} \tag{11}$$

where, u and v respectively denote the horizontal and vertical range of the SSM. After that, a surround suppression response $C_{\alpha,\sigma}(x, y)$ was designed as below

$$C_{\alpha,\sigma}(x, y) = H(C_\sigma(x, y) - \alpha t_\sigma(x, y)) \tag{12}$$

where, $C_\sigma(x, y)$ is the corner magnitude, $t_\sigma(x, y)$ is the suppression term, and $H(z) = z$ when $z \geq 0$ and $H(z) = 0$ when $z < 0$, and α controls the strength of the surround suppression. Therefore, unwanted points of surrounding are suppressed. As a consequence, the STIPs detected by this algorithm are more repeatable and stable, and background points are suppressed by using the temporal constraint. The algorithm can detect effective STIPs in motion regions of video and can also handle complex situations, *e.g.*, illumination variation, perspective transformation, background confusion and camera movement. Its time complexity mainly depends on the size of the input video. However, in some special cases, *e.g.,* the combination of all above-mentioned complex situations, a large number of STIPs detected by the selective STIP detector are background points.

Additionally, the V-FAST detector which is extended from the fast corner detector [36] was proposed by Yu *et al.* [37]. It is a real-time solution using the local appearance and structural information of moving objects to detect a dense set of STIPs.

B. STB AND ITS DETECTION ALGORITHMS

Spatio-temporal blobs refer to regions with brightness difference or color difference compared to the surrounding regions in the video volume. The essence of spatio-temporal blob detection is to calculate the extreme values of the second-order derivative. In the past years, researchers have devoted themselves to search for effective differential operators and proposed fast blob detection algorithms to improve the detection efficiency.

One kind of the STB detectors is based on the three-dimensional Laplace of Gaussian (3DLoG) extended from Laplace of Gaussian (LoG). The 3DLoG operator is defined as below

$$\begin{aligned} LoG(x, y, t; \sigma, \tau) &= \sigma^3 \nabla^2 \mathrm{g} \\ &= \frac{x^2 + \mathrm{y}^2 + \mathrm{t}^2 - 3\sigma\tau}{2\pi\tau^3} e^{-\frac{x^2+y^2+t^2}{2\sigma\tau}} \end{aligned} \tag{13}$$

where, g is 3D Gaussian kernel with the spatial and temporal scale parameters respectively denoted as σ and τ, and $\nabla^2 g$ is the Laplace operator of g.

3D Difference of Gaussian (3DDoG) operator constructed by subtracting the adjacent Gaussian scales is an approximation of 3DLoG operator. Based on 3DDoG operator, 3D Scale Invariant Feature Transform (3DSIFT) detector was proposed [38]–[40]. 3DSIFT feature has advantages of rotation invariance, illumination variation, and robustness to noise. However, similar to Harris3D detector, 3DSIFT detector extended from SIFT detector [2] processes the temporal information by using the method of processing the spatial information. As a consequence, the detector can only captures the implicit motion information.

Aiming at solving the shortages of 3DSIFT, Al Ghamdi *et al.* [41] proposed spatio-temporal SIFT (ST-SIFT) detector. Unlike 3DSIFT detector where 3D spatial pyramids are constructed, ST-SIFT detector takes a video as a sequence of image frames and changes z axis of 3DSIFT detector to time axis. Firstly, multi-levels of the volume pyramid of Gaussian L are constructed with spatially and temporally downsampling, and incremental convolution. The spatially and temporally downsampling refers that the lower level is spatially and temporally downsampled from the upper level. The incremental convolution means that the convolution of each level with 3D Gaussian kernel G is calculated to creat a Guassian pyramid as shown in Eq.(14).

$$L(x, y, t; \sigma, \tau) = G(x, y, t; \sigma, \tau) * I(x, y, t; \sigma, \tau) \quad (14)$$

Then, for each level in the Gaussian pyramid, the spatio-temporal DOG pyramid is generated as below

$$\begin{aligned} D(x, y, \sigma, \tau) &= (G(x, y, K\sigma, \tau) - G(x, y, \sigma, \tau)) * I(x, y, \sigma, \tau) \\ &= L(x, y, K\sigma, \tau) - L(x, y, \sigma, \tau) \end{aligned} \quad (15)$$

The essence of ST-SIFT detector is that the video volume is sliced into three planes, i.e., *xy*, *yt*, and *xt*, and DoG pyramids for these three slices are generated respectively. Searching the local maximum or minimum in each slice pyramid, and extreme points found in all three slice pyramids are selected as STIPs. ST-SIFT detector is a spatio-temporal extension of SIFT detector [2]. It performs well at representing events in video because STIPs detected by ST-SIFT detector reflect the texture information of the spatial domain and the motion information of the temporal domain. Therefor, it can clearly distinguish the similar behaviors, such as walking and running. Moreover, it is robust to resist scale variation and orientation variation for the use of SIFT features. However, ST-SIFT detector mixes the spatial axis and the temporal axis to form *xt* and *yt* planes without accounting the differences between the space and time, which is similar to 3DSIFT detector at some extent.

Guo [42] proposed s-t SIFT detector with the purpose of solving the disturbance of noises in video with complex scene. This detector divides the spatio-temporal domain into two independent subspaces, i.e., the spatial domain and the temporal domain, and searches the maximum of DoG in the spatial domain to obtain spatial interest points. Then, it uses a temporal pyramid to filter spatial interest points and reserve those points with extreme values in the temporal domain simultaneously. Obviously, those points eventually reserved are spatio-temporal SIFT points. This detector has many excellent characteristics for video with complex scene, including scale invariance, illumination invariance and anti-interference. And s-t SIFT detector can perform well at locating those points with extreme value in the scale domain. Not only does it keep the directionality of STIPs but also achieves rotation invariance while locating extreme points in scale space. However, s-t SIFT detector is time-consuming and the computation speed is slower than that of Harris3D detector [12], so it obviously cannot achieve real-time performance. Meanwhile, the detected STIPs are sparse because this detector independently uses the temporal pyramid to filter spatial interest points, which ignores the spatio-temporal correlation of video pixels in the video volume.

Chen and Hauptmann [43] proposed an algorithm called MoSIFT detector, handling the spatial domain and the temporal domain separately. It uses SIFT detector [2] to detects SIFT points at each frame of video as candidate STIPs. At the same time, optical flow pyramids are constructed to compute optical flow, which detects the movements of regions in video. Only candidate STIPs which contain sufficient motion in the optical flow pyramids can be defined as real STIPs. Its good performance mainly due to the introduction of the well-known optical flow approach. It can be used in real-world video which contains highly crowed scenes, severely cluttered background and large viewpoint variation. However, it is scale invariant in the spatial domain but not scale invariant in the temporal domain.

Another kind of STB detectors is based on Hessian matrix, the matrix composed by the second-order derivatives of scale space L [4], [14], [44]. 3D Hessian matrix is a spatio-temporal extension of Hessian matrix proposed by Willems *et al.* [14], which is shown as Eq.(16).

$$H = \begin{bmatrix} L_{\mathrm{xx}} & L_{\mathrm{xy}} & L_{\mathrm{xt}} \\ L_{\mathrm{yx}} & L_{\mathrm{yy}} & L_{\mathrm{yt}} \\ L_{\mathrm{tx}} & L_{\mathrm{ty}} & L_{\mathrm{tt}} \end{bmatrix} \quad (16)$$

Willems developed Hes-STIP detector that uses approximative box-filter operations on an integral video to obtain multi-scale space. Firstly, a input video is transformed to an integral video and all 3D Gaussian second-order derivatives are approximated by 3D box-filter equivalents. Then, the strength of each interest point is computed at a certain scale, which is defined as Eq.(17).

$$S = |det(H)| = L_{xx}L_{yy}L_{tt} \quad (17)$$

After that, they apply non-maximum suppression over the space $(x, y, t; \sigma, \tau)$ and STIPs with the local maximum of S are obtained. To achieve scale selection, they use the γ-normalized determinant to search STIPs.

Compared to the iterative approach of Harris3D detector, Hes-STIP detector removes the need for an iterative scheme to avoid the divergence problem in the progress of calculation. Its computation speed is significantly improved owing

TABLE 1. Comparison between spatio-temporal interest point detection algorithms.

Algorithm	Feature Sets	Scale Invariance	Execution Rate	Applicable Scene	Camera Movement	Cluttered Background
Harris3D [12]	sparse	×	slow	simple	×	√
Cuboid [13]	dense	×	fast	simple	×	×
Ke [31]	dense	√	real-time	complex	√	√
Oikonomopoulos [32]	sparse	√	slow	simple	×	×
Laptev [33]	sparse	√	--	complex	√	√
Shabani [15]	dense	√	fast	complex	√	√
Liu [34]	dense	√	slow	complex	√	√
Selective STIP [16]	dense	√	--	complex	√	√
V-FAST [36]	dense	√	real-time	complex	√	√
ST-SIFT [41]	dense	√	--	complex	--	--
s-t SIFT [42]	sparse	√	slow	complex	--	√
MoSIFT [43]	dense	√	--	complex	√	√
Hes-STIP [14]	dense	√	fast	simple	√	×
MVSTIPs [45]	dense	√	slow	complex	√	√

to the contribution of the integral video and the box-filters. The scale space can be efficiently constructed by upscaling the box-filters and STIPs are scale invariant in the spatio-temporal domain. However, not all eigenvalues of H have identical signs, which means that not only blobs but also saddle points are detected. Consequently, most of the detected STIPs belong to the background in the case of dynamic backgrounds and the camera movement.

In 2014, Li *et al.* [45] presented a robust method based on multi-velocity spatio-temporal interest points (MVSTIPs) for human action recognition. They firstly use multi-direction motion energy (ME) filters [46] at different speeds to detect significant changes of video pixels in both the spatial domain and the temporal domain. ME filter is designed as Eq.(18).

$$E_{v,\theta}(x,y,t)=\sqrt{r^2_{v,\theta,0}(x,y,t)+r^2_{v,\theta,\pi/2}(x,y,t)} \quad (18)$$

where, $r_{v,\theta,0}(x,y,t)$ and $r_{v,\theta,\pi/2}(x,y,t)$ are 3D Gabor responses with a phase difference of $\pi/2$, and they are computed by convolving the signal $l(x,y,t)$ with the 3D Gabor kernel $g_{v,\theta,\phi}(x,y,t)$ as Eq.(19).

$$r_{v,\theta,\phi}(x,y,t)=l(x,y,t)*g_{v,\theta,\phi}(x,y,t) \quad (19)$$

Then, a surround suppression model [46] is used to rectify the deviation of ME filters caused by the camera movement and complex backgrounds. Finally, local maximum filters are utilized to obtain MVSTIPs at different speeds.

In the process of the detection, ME filters makes great contribution to detect STIPs because they are sensitive to motions at multi-speeds in various directions in video. However, ME filters respond strongly to texture areas and complex backgrounds, resulting in inaccurate detection. And the surround suppression model is introduced to reserve ME responses around the cubic center but suppress ME responses in farther surrounds, so the surround suppression model effectively reduces the false and redundant points caused by complex and dynamic backgrounds [20]. MVSTIP detector obtains an amount of STIPs with sufficient motion information and achieves a high recognition accuracy in real-world video.

III. SUMMARY AND CHALLENGES OF STIP DETECTION

A. SUMMARY

We have summarized the implementation of various STIP detection algorithms including STC and STB detection algorithms. Also, the advantages and limitations of these algorithms have been presented.

Table 1 shown as follows summarizes the most important properties of the detection algorithms mentioned above. We compare many existing STIP detection algorithms in 6 aspects, including feature sets, scale invariance, execution rate, applicable scene, camera movement and cluttered background. The term 'simple' denotes that the algorithm merely performs in scenes with situations of single action, fixed perspective or constant illumination. The term 'complex' denotes the scene including serious shadow, perspective occlusion, illumination changes or crowded behaviors. The symbol '$\surd$' means that the algorithm has the corresponding properties and '$\times$' represents that the algorithm does not have the corresponding properties.

B. CHALLENGES

Video are characterized by large amount of data which are diverse and unstructured. And there exist many types of uncertain interference factors in real-world video, *e.g.*, illumination change, background clutter, perspective occlusion, perspective transformation and low resolution. Consequently, STIP detection algorithms play an important role in each STIP-based approach. The research challenges of the existing STIP detection algorithms can be summarized as follows.

1) CONSTRUCTING A STIP DETECTION MODEL WHICH CAN SYNTHETICALLY ANALYSE THE SPATIAL AND TEMPORAL INFORMATION OF VIDEO IS ONE OF THE CHALLENGES

The signal of video is causal in the temporal domain but non-causal in the spatial domain. Moreover, the temporal signal is different from the spatial signal in both the dimension and resolution. Therefore, the detector simply expanded from the detector of 2D image features ignores the differences between the temporal signal and the spatial signal. Most of

STIP detection algorithms split the correlation between the spatial domain and the temporal domain, without considering the spatio-temporal correlation of video pixels in the video volume. Such as Harris3D detector separately searches for local extreme points in the spatio-temporal domain [12], and the Cuboid detector separately uses Gaussian filters in the spatial domain and Gabor filters in the temporal domain [13]. These will lead to the detected STIPs are sparse and not robust.

2) CONSIDERING BOTH THE QUANTITY AND THE EFFICIENCY OF STIP DETECTION ALGORITHMS, ESPECIALLY IN THE CASE OF ONLY SMOOTH MOVEMENT

There is a conflict between the quantity and the efficiency of the detected STIPs in existing algorithms. For instance, the STIPs detected by algorithm [32], algorithm [33] and s-t SIFT [42] are of good quality but sparse. On the contrary, Cuboid detector detect dense STIPs, but a mass of invalid feature points are contained [13]. In particular, state-of-the-art STIP detection algorithms are not suitable for smooth movement. Because STIP detection algorithms detect those points with significant local variation of the pixel intensity in the spatio-temporal domain. They can detect STIPs from moving objects with constantly changing directions. However, for smooth movement, such as unidirectional uniform movements and uniformly accelerated movements, STIP detection algorithms cannot obtain a good response and is difficult to accurately reflect the actual motion characteristics from detection results.

3) REAL TIME IS A RESEARCH CHALLENGE OF STIP DETECTION ALGORITHMS AND IT IS THE KEY OF THESE ALGORITHMS CAN BE WILDLY USED

It is important of detecting features in real time for the analysis of real-world video. However, most of the existing algorithms achieve detecting large number and high accurate features, but they may be found the execution rate are slow and cannot satisfy the requirement of real time [14], [31], [37]. It is mainly due to the large computational cost of constructing the scale space and searching STIPs. These algorithms cannot be applied in the applications which require real time. Therefore, it is essential to develop STIP detection algorithms with high efficiency.

4) EFFECTIVELY DETECTING STIPS FOR REAL-WORLD VIDEO IS STILL CHALLENGING

Most of STIP detection algorithms are ineffective for real-world video which contains background noises, perspective transformation,and serious shadows of moving objects. Moreover, due to unknown ego disturbances such as the camera moving or shaking, background noises and perspective transformation are produced and the motion information of video is distorted to some extent [33], which will lead to STIP detection algorithms detecting false STIPs. And serious shadows will be wrongly regarded as objects themselves and their features are mistakenly detected, this is due to serious shadows are distinct from their backgrounds but moving with objects. In addition, STIP detection is also affected by the occlusion, illumination changes and crowded behaviors. Therefore, these interference factors make STIP detection algorithms obtain a large number of background noises and false STIPs. In the result, the efficiency of STIP detection becomes low and intelligent video analysis becomes more and more difficult.

IV. APPLICATIONS OF STIP

In recent years, intelligent video analysis based on STIP has become a hot topic in computer vision and is now widely applied in human action recognition [17]–[20], video surveillance [21]–[23], video summarization [24]–[26] and intelligent video retrieval [27]–[29].

A. HUMAN ACTION RECOGNITION

Human action recognition is a significant application of STIP in video. Firstly, STIPs of video are detected and calculated their descriptors. Thereafter, codebooks [47]–[49] are applied to encode STIPs by using vector quantization algorithms. Finally, pattern classification algorithms, such as Support Vector Machine (SVM), k-Nearest Neighbor (KNN) and Bayes classifier, are used to identify human actions. Human action recognition algorithms based on STIP have been developed from the early stages of recognizing simple and single actions, and now are applied to recognize complex actions [17], [18] and crowd behaviors [19], [20] in real-world video. Based on the application in human action recognition, STIPs have been applied in many other applications, such as virtual reality [50]–[54], motion analysis [5], [21], [54] and human-computer interaction [17], [33], [36].

B. ANOMALY DETECTION IN VIDEO SURVEILLANCE

STIP plays an important role in video surveillance. An abnormal event detection algorithm based on STIP is introduced in literature [21] to apply to traffic video surveillance. The location distribution and the velocity distribution of extracted STIPs are used to classify abnormal behaviors in video. Algorithm [22] detects STIPs that represent abnormal events in video at different levels of spatio-temporal contexts. Moreover, a new traffic anomaly detection algorithm based on STIP is introduced to detect the traffic anomaly on intersection traffic video and main road video [23].

C. VIDEO SUMMARIZATION

Using STIP detection algorithms to extract the key information of video facilitates the development of the video content summarization technology. A video summarization method based on STIP was proposed in literature [24]. It utilizes the spatio-temporal Hessian matrix to detect STIPs. Then it uses these detected STIPs to obtain the candidate video summarization, and the final video summarization is obtained through eliminating the clapperboard images from redundant video fragments. Literature [25] presented a video summarization approach based on machine learning, which extracts STIPs by comparing two adjacent frames. Key video frames are obtained because the detected STIPs contain important information of video. Furthermore, literature [26] proposed a framework for detecting static points and STIPs, which can automatically and quickly detect moving objects in video and be applied to video summarization.

D. CONTENT-BASED VIDEO RETRIEVAL

Content-based video retrieval system detects STIPs from massive videos and then returns those videos required from users. Literature [27] proposed a spatio-temporal salient objects-based method, which detects STIPs integrated with spatial characteristics and the motion information. It is in favor of analyzing the motion of video. Moreover, Enhanced Multi-Spectro-Temporal Curvature Scale Space (EMST-CSS), a novel feature representation, was proposed in paper [28]. The algorithm captures STIPs and performs effectively for content-based video retrieval. Moreover, article [29] realized content-based video retrieval by using STIPs of the spatio-temporal volume to represent the dynamic of geometric features.

V. PROSPECTS

Video analysis and processing has been widely applied in the field of computer vision. As an important part of video analysis and processing, STIP detection has attracted many researchers' attention. However, there are many challenges mentioned above needed to be solved. Therefore, STIPs detection will still be a hot research topic in the future, and there are some research prospects in the future are shown as follows.

1) ENHANCING THE SCENE ADAPTABILITY OF STIP DETECTION ALGORITHMS

Most of the researches of STIP detection are based on the standard video datasets. Video with single or group behavior, static or dynamic background, fixed or moving camera have been strictly classified. To some extent, experimenting with standard datasets avoids the complexity of the research and ensures the robustness in STIP extraction. However, it fundamentally limits these algorithms to apply to real-world video. The future research should be wildly applied to real-world video and could effectively solve the challenges of illumination change, background clutter, camera movement, perspective occlusion, perspective transformation and low resolution.

2) IMPROVING THE EFFICIENCY OF STIP DETECTION ALGORITHMS TO ACHIEVE REAL-TIME PERFORMANCE

Video analysis and processing based on STIP has been wildly applied in the fields of unsupervised intelligent video surveillance, human-computer interaction and human action recognition. These applications are required to detect STIPs in real time, so the efficiency of STIP detection algorithms must be improved to achieve real-time performance. On the one hand, STIP detection algorithms must be optimized. On the other hand, GPU programming and distributed computation can be used in algorithms to improve the operating speed.

3) SOLVING THE BIG DATA PROBLEMS OF VIDEO

With the advent of the era of big data, video data is regarded as the main component of big data. Massive high-definition videos become an essential part of our daily life. However, in the process of analyzing and processing high-definition video, it faces the problems caused by large amounts of data, high processing speed and various data types, which need to be addressed urgently.

4) IMPROVING THE FEATURE DETECTION MODEL

Most of the existing STIP detection algorithms are spatio-temporal extensions of 2D image detection algorithms. They split the correlation between the spatial domain and the temporal domain, ignoring the uncertainty of temporal correlation [5]. It is the result that the detected STIPs cannot achieve robustness and uniqueness. Therefore, only taking spatio-temporal correlation into account and improving novel feature detection models could push to produce more effective STIP detection algorithms for video analysis and processing.

5) COMBINING STIPs WITH DEEP LEARNING METHOD TO ANALYZE AND PROCESS VIDEO

Deep learning is a major breakthrough in the field of machine learning, becoming a hot topic of research nowadays. However, deep learning in the field of video is still in its infancy due to the high computational complexity and the large scale of learning. Bruna and Mallat [55] proposed a deep convolution network whose first layer outputs SIFT-type descriptors, which has a good performance in handwritten digit recognition. Therefore, it is feasible to combine STIPs with deep learning, *e.g.*, STIPs can be put into the deep neural network to obtain more features and get higher recognition accuracy on video recognition and classification.

VI. CONCLUSION

In conclusion, this paper presents a comprehensive review on STIP detection algorithms. The classical and current popular STIP detection algorithms have been summarized, with the analysis of their advantages and shortcomings. Also, the challenges of various algorithms proposed by many researchers have been pointed out. The STIP-based detection is wildly applied in many applications. We believe that with continuous developments in computer vision research field and the progress of computer hardware performance, the existing problems can be solved step by step and STIP detection of video will be widely applied to improve the quality of human life.

REFERENCES

[1] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, vol. 15, 1988, pp. 5210–5244.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[3] L. Zhang, H. Wang, T. Deng, and X. He, "Improving integrality of detected moving objects based on image matting," *Chin. J. Electron.*, vol. 23, no. 4, pp. 742–746, 2014.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[5] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[6] K. Mikolajczyk, B. Leibe, and B. Schiele, "Local features for object class recognition," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct. 2005, pp. 1792–1799.

[7] L. Wang, W. Xie, and J. Pei, "Patch-based dark channel prior dehazing for RS multi-spectral image," *Chin. J. Electron.*, vol. 24, no. 3, pp. 573–578, 2015.

[8] X. Li, J. Shi, Y. Dong, and D. Tao, "A survey on scene image classification," *Scientia Sinica Inf.*, vol. 45, no. 7, p. 827, 2015.

[9] L. Liu, L. Shao, X. Li, and K. Lu, "Learning spatio-temporal representations for action recognition: A genetic programming approach," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 158–170, Jan. 2016.

[10] Y. Li, W. Liu, X. Li, Q. Huang, and X. Li, "GA-SIFT: A new scale invariant feature transform for multispectral image using geometric algebra," *Inf. Sci.*, vol. 281, pp. 559–572, Oct. 2014.

[11] X. Zhen, L. Shao, and X. Li, "Action recognition by spatio-temporal oriented energies," *Inf. Sci.*, vol. 281, pp. 295–309, Oct. 2014.

[12] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.

[13] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Jun. 2005, pp. 65–72.

[14] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 650–663.

[15] H. Shabani, D. A. Clausi, and J. S. Zelek, "Towards a robust spatio-temporal interest point detection for human action recognition," in *Proc. Can. Conf. IEEE Comput. Robot Vis. (CRV)*, Dec. 2009, pp. 237–243.

[16] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez, "Selective spatio-temporal interest points," *Comput. Vis. Image Understand.*, vol. 116, no. 3, pp. 396–410, Mar. 2012.

[17] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2046–2053.

[18] H. Zhang and L. E. Parker, "4-dimensional local spatio-temporal features for human activity recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, Jun. 2011, pp. 2044–2049.

[19] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1446–1453.

[20] C. Wang and C. Dong, "Spatial-temporal words learning for crowd behavior recognition," *Int. J. Sci. Eng. Invest.*, vol. 1, no. 3, pp. 1–6, 2012.

[21] L. Cui, K. Li, J. Chen, and Z. Li, "Abnormal event detection in traffic video surveillance based on local features," in *Proc. 4th Int. Congr. Image Signal Process. (CISP)*, vol. 1, Oct. 2011, pp. 362–366.

[22] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Comput. Vis. Image Understand.*, vol. 115, no. 3, pp. 323–333, 2011.

[23] Y. Li, W. Liu, and Q. Huang, "Traffic anomaly detection based on image descriptor in videos," *Multimedia Tools Appl.*, vol. 75, no. 5, pp. 2487–2505, 2016.

[24] R. Laganiére, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu, "Video summarization from spatio-temporal features," in *Proc. ACM Workshop Video Summarization*, Vancouver, BC, Canada, Oct. 2008, pp. 144–148.

[25] W. Ren and Y. Zhu, "A video summarization approach based on machine learning," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process. (IIHMSP)*, Apr. 2008, pp. 450–453.

[26] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, p. 15, 2009.

[27] H. P. Gao and Z. Q. Yang, "Content based video retrieval using spatiotemporal salient objects," in *Proc. Int. Symp. Intell. Inf. Process. Trusted Comput. (IPTC)*, Sep. 2010, pp. 689–692.

[28] C. Chattopadhyay and S. Das, "Enhancing the MST-CSS representation using robust geometric features, for efficient content based video retrieval (CBVR)," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, May 2012, pp. 352–355.

[29] C. Chattopadhyay and A. K. Maurya, "Multivariate time series modeling of geometric features of spatio-temporal volumes for content based video retrieval," *Int. J. Multimedia Inf. Retr.*, vol. 3, no. 1, pp. 15–28, 2014.

[30] L. M. Florack, B. M. T. H. Romeny, J. J. Koenderink, and M. A. Viergever, "Scale and the differential structure of images," *Image Vis. Comput.*, vol. 10, no. 6, pp. 376–388, 1992.

[31] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, May 2005, pp. 166–173.

[32] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 3, pp. 710–719, Jun. 2005.

[33] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Comput. Vis. Image Understand.*, vol. 108, no. 3, pp. 207–229, 2007.

[34] C. Liu, Y. Chen, and M. Wang, "Spatio-temporal interest point detection in cluttered backgrounds with camera movements," *J. Image Graph.*, vol. 18, no. 8, pp. 982–989, 2013.

[35] J. Weickert, "A review of nonlinear diffusion filtering," in *Scale-Space Theory in Computer Vision*. Berlin, Germany: Springer, 1997, pp. 1–28.

[36] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," *Computer Vision-ECCV*. Berlin, Germany: Springer, 2006, pp. 430–443.

[37] T. H. Yu, T. K. Kim, and R. Cipolla, "Real-time action recognition by spatiotemporal semantic and structural forests," *BMVC*, vol. 2, no. 5, p. 6, 2010.

[38] W. Cheung and G. Hamarneh, "N-sift: N-dimensional scale invariant feature transform for matching medical images," in *Proc. 4th IEEE Int. Symp. Biomed. Imag. Nano Macro.*, Jun. 2007, pp. 720–723.

[39] S. Allaire, J. J. Kim, S. L. Breen, D. A. Jaffray, and V. Pekar, "Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2008, pp. 1–8.

[40] D. Ni *et al.*, "Reconstruction of volumetric ultrasound panorama based on improved 3D SIFT," *Comput. Med. Imag. Graph.*, vol. 33, no. 7, pp. 559–566, 2009.

[41] M. Al Ghamdi, L. Zhang, and Y. Gotoh, "Spatio-temporal SIFT and its application to human action classification," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 301–310.

[42] Y. Guo, "Spatio-temporal SIFT interest points detection in videos," Ph.D dissertation, College Comput. Sci. Technol., Zhejiang Univ., Hangzhou, China, 2009.

[43] M. Y. Chen and A. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos," *Ann. Pharmacotherapy*, vol. 39, no. 1, pp. 150–152, 2009.

[44] J. Liu and M. Shah, "Learning human actions via information maximization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Jun. 2008, pp. 1–8.

[45] C. Li, B. Su, J. Wang, H. Wang, and Q. Zhang, "Human action recognition using multi-velocity STIPs and motion energy orientation histogram," *J. Inf. Sci. Eng.*, vol. 30, no. 2, pp. 295–312, 2014.

[46] N. Petkov and E. Subramanian, "Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal Gabor filters with surround inhibition," *Biol. Cybern.*, vol. 97, nos. 5–6, pp. 423–439, 2007.

[47] Y. Li, Q. Huang, W. Xie, and X. Li, "A novel visual codebook model based on fuzzy geometry for large-scale image classificatio," *Pattern Recognit.*, vol. 48, no. 10, pp. 3125–3134, 2015.

[48] Y. Li, W. Liu, Q. Huang, and X. Li, "Fuzzy bag of words for social image description," *Multimedia Tools Appl.*, vol. 75, no. 3, pp. 1371–1390, 2016.

[49] L. Zhang, M. Wang, R. Hong, B. Yin, and X. Li, "Large-scale aerial image categorization using a multitask topological codebook," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 535–545, Feb. 2016.

[50] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Nov. 2005, pp. 984–989.

[51] Z. Lin, Z. Jiang, and L. S. Davis, "Prototype-based methods for human movement modeling," in *Computer Vision*. New York. NY, USA: Springer, 2014, pp. 651–654.

[52] W. Zhou, H. Hu, and P. J. Smith, "Recent advances in camera tracking and virtual reality," in *Proc. Inf. Technol. Proc. Int. Symp. Inf. Technol. (ISIT)*, Dalian, China, Oct. 2014, pp. 14–16.

[53] A. Chaudhary, J. L. Raheja, K. Das, and S. Raheja. (Mar. 2013). "Intelligent approaches to interact with machines using hand gesture recognition in natural way: A survey." [Online]. Available: https://arxiv.org/abs/1303.2292

[54] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view human action recognition: A survey," in *Proc. 9th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Oct. 2013, pp. 522–525.

[55] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886, Aug. 2013.

YANSHAN LI received the M.Sc. degree from the Zhejiang University of Technology in 2005, and the Ph.D. degree from the South China University of Technology, China. He is currently an Associate Professor with the ATR National Key Laboratory of Defense Technology, Shenzhen University, China. His research interest covers computer vision, machine learning, and image analysis.

RONGJIE XIA received the B.E. degree in information and engineering from Shenzhen University, Shenzhen, China, in 2017, where he is currently pursuing the M.S. degree in signal and information processing with the ATR National Key Laboratory of Defense Technology. His research interests include intelligent information processing, video processing, and pattern recognition.

QINGHUA HUANG received the Ph.D. degree in biomedical engineering from the Hong Kong Polytechnic University, Hong Kong, in 2007. He joined the School of Electronic and Information Engineering, South China University of Technology, China, in 2008. He is currently a Full Professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), School of Electronics and Information, Northwestern Polytechnical University, China. He has been a Guest Distinguished Professor with the College of Information Engineering, Shenzhen University, Shenzhen, China, since 2016. He has published 100+ technical papers, and serves as a Guest Editor, an Editorial Board Member, or a Reviewer for more than 40 international journals. His research interests include pattern recognition, medical imaging, bioinformatics, intelligent computation, and its applications.

WEIXIN XIE received the degree from Xidian University, Xi'an. He was a Faculty Member with Xidian University in 1965. From 1981 to 1983, he was a Visiting Scholar at the University of Pennsylvania, USA. In 1989, he was invited to the University of Pennsylvania, as a Visiting Professor. He is currently with the School of Information Engineering, Shenzhen University, China. His research interests include intelligent information processing, fuzzy information processing, image processing, and pattern recognition.

XUELONG LI is currently a Researcher (Full Professor) with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi, China.