Video Saliency Map Detection by Dominant Camera Motion Removal

Chun-Rong Huang, *Member, IEEE,* Yun-Jung Chang, Zhi-Xiang Yang, Yen-Yu Lin*, *Member, IEEE*

***Abstract*—We present a trajectory based approach to detect salient regions in videos by dominant camera motion removal. Our approach is designed in a general way so that it can be applied to videos taken by either stationary or moving cameras without any prior information. Moreover, multiple salient regions of different temporal lengths can also be detected. To this end, we extract a set of spatially and temporally coherent trajectories of keypoints in a video. Then, velocity and acceleration entropies are proposed to represent the trajectories. In this way, long-term object motions are exploited to filter out short-term noises, and object motions of various temporal lengths can be represented in the same way. On the other hand, we are inspired by the observation that the trajectories in backgrounds, i.e., the non-salient trajectories, are usually consistent with the dominant camera motion no matter whether the camera is stationary or not. We make use of this property to develop a unified approach to saliency generation for both stationary and moving cameras. Specifically, *one-class SVM* is employed to remove the consistent trajectories in motion. It follows that the salient regions could be highlighted by applying a diffusion process to the remaining trajectories. In addition, we create a set of manually annotated ground truth on the collected videos. The annotated videos are then used for performance evaluation and comparison. The promising results on various types of videos demonstrate the effectiveness and great applicability of our approach.**

***Index Terms*—Video saliency map, trajectory, one-class SVM**

I. Introduction

The human visual system (HVS) perceives the world and provides visual information for human beings. It is known that only a small fraction of the observable area is critical for humans to understand and interpret the world. Hence, recognizing saliency maps, which record the distribution of human's attention, is helpful in understanding how HVS works and analyzing the content of images/videos. Thus, it becomes one of the most important research issues in the fields of computer vision, image and video processing. As a key component to image/video understanding, saliency maps can be used in a wide range of applications, such as object detection, image retargeting, and compression.

While most research effort has been made on saliency detection in still images, one of the research trends has been shifted to discover saliency maps in videos. Comparing with images, videos bring richer visual evidences, such as the structural information and the motion patterns of objects. These evidences help in object localization and noise removal, and greatly benefit the task of saliency generation. However, detecting saliency maps in videos is challenging owing to both intrinsic and extrinsic factors. For example, videos can be taken by either stationary or moving cameras, and objects in videos may mutually occlude and appear with various temporal lengths. The yielded variations of the appearances in salient regions also complicate saliency map detection. Furthermore, the large volume of video data in general cause inefficiency in processing. It hinders the applicability of saliency maps in real-time tasks.

In this work, we aim at developing a general and efficient algorithm for video saliency map generation. We capture long-term object motions to generate video saliency maps, and develop a unified algorithm that can work on videos taken by both stationary and moving cameras without any prior information. Our approach can distinguish itself with the following two contributions.

First, to solve the observation length problem, we track salient keypoints in videos to yield *trajectories*. The trajectories on objects naturally identify the observation lengths of the corresponding objects. A compact and effective descriptor based on the velocity and acceleration entropies is designed to characterize the trajectories. In this manner, the task of saliency detection is formulated as a binary classification problem over trajectories, i.e., saliency vs. non-saliency.

Second, we utilize one-class SVM to perform the unsupervised classification. Motivated by the observation that most trajectories in backgrounds are consistently enclosed by the dominant camera motion, one-class SVM, which separates alike data from the outliers, is applied to removing non-salient trajectories. As a result, our approach can work on videos captured by both stationary and moving cameras without knowing the dominant camera motion.

In addition, we selected a few videos, and manually annotated the salient regions in the videos. These videos were taken by either stationary or moving cameras. Among them, one or multiple salient regions of different temporal lengths are included. They span a wide spectrum of practical conditions. We will make the annotated ground truth publicly available, and believe that it will be a good resource for performance evaluation of video-based saliency detection methods.

This work is supported in part by grants NSC 101-2221-E-005-086-MY3 and NSC 102-2221-E-001-025.

C.-R. Huang, Y.-J. Chang, and Z.-X. Yang are with the Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan (e-mail: crhuang@nchu.edu.tw).

Y.-Y. Lin, Y.-J. Chang, and Z.-X. Yang are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan (e-mail: yylin@citi.sinica.edu.tw).

II. Related Work

In this section, we firstly review a few representative image-based saliency models. Subsequently, some motion-aware approaches to video saliency detection are discussed.

A. Image-based Visual Saliency

One of the pioneering saliency models was proposed by Itti et al. [1]. They computed low-level features, such as color, intensity and orientations, to retrieve the basic elements of *contrast*, which serve as the clues for highlighting the most attractive regions in an image. Methods of attractive region detection can be roughly divided into two categories. The first category contains methods developed based on the *center-surround hypothesis* [1], [2], [3], [4], [5], [6], [7], which attributes the saliency as the distinctiveness of an image region over its surroundings. Methods of the second categories, such as [8], [9], instead treat the distinctiveness of certain features in a global fashion. These two manners are complementary, and can also be jointly utilized in terms of local and global patches to boost the detection results. However, Perazzi et al. [10] asserted that the contributions of individual features still remain obscure. They hence proposed to decompose an image into perceptually homogeneous elements, and redefined the contrast in terms of uniqueness and the spatial distribution of those elements.

Apart from using low-level image features as stimulus signals in a *bottom-up* scheme, *top-down* approaches, e.g., [2], [11], [8], [12], investigate learned class-dependent features for specific pattern discovery. These approaches recognize the most salient regions, which can distinguish themselves from the rest, in a class-specific manner. Besides, Borji [12] further integrated the low-level features with the top-down cognitive visual features, such as the outcomes of object detectors. Both bottom-up and top-down visual evidences are then taken into account. Top-down approaches have their restrictions, because they are applicable by assuming that prior knowledge about the salient regions is given. However, prior knowledge is not generally available in practice.

In a bottom-up framework, images can be characterized in diverse ways. For example, the information from image-level spectrums in [13] is directly analyzed instead of extracting features based on certain prior knowledge of the targets. In most cases, using multiple kinds of features to jointly describe images is appreciated, because salient regions are more likely to be distinguished from the rest by at least one type of features. It follows that the contrast in the salient regions yields to draw human attentions.

Among off-the-shelf features, color-based features are quite important, because they closely relate to human perception. They are thus adopted by many works, e.g., [1], [7], [14], [15]. However, most conventional color histograms contain limited information. For further enhancement, the *color co-occurrence histogram* was introduced in [15] to extend traditional 1D color histograms to 2D ones, in which not only the color distributions but also their spatial configuration are recorded. Furthermore, Borji and Itti [14] jointly employed RGB and Lab color spaces to lead to more clues for improving saliency detection.

When video sequences are considered, a naïve way of saliency map detection is to directly apply one of the image-based methods, such as [7], to video frames independently. However, human visual systems determine salient regions in still images and video sequences by different ways. It has been pointed out in the study [16] that human attentions are attracted by motions of video content between adjacent frames. Consequently, the motions can provide strong evidences for identifying saliency maps in videos [17]. In contrast, methods that are developed for still images often result in incoherent saliency maps because they fail to take the temporal information between adjacent frames into account.

B. Motion-aware Visual Attention

Compared with spatial information, temporal motions can also provide abundant information particularly for video saliency map detection. However, the analysis of motion features is challenging, since motions of salient objects and those of backgrounds may be complicated and even mix together. In the following sections, we review some video-based saliency map detection methods with a stationary camera and a moving camera, respectively.

1) Stationary Camera: When the camera is static, all the present motions can potentially attract human attentions, and this assumption has been extensively exploited in saliency map detection, e.g., [18], [19], [20], [21], [22], [23]. Without background motions, object motions as well as salient regions can typically be identified based on two-frame differences, e.g., [6], [23], [18], [19], [20], [24]. These approaches make use of spatiotemporal cues for determining the salient regions that are more consistent with human's first sight.

However, there are still unsolved issues for approaches of this category. First, the balance between spatial and temporal information may not be deterministic [23], because it is most likely content-dependent. Second, these approaches may suffer from the problem caused by short-term noise, and yield incoherent saliency maps in successive frames. This is because the salient regions induced by motions from merely two frames are not sufficient to stably represent the saliency of the entire video. As a solution, the temporal saliency map in [25] is computed based on homographies of several frames. Yet the assumption of planar motions, or homography, in videos is too general to be true in practice, even though the long-term temporal information rather than two-frame differences is considered.

Motion pattern discovery over a certain range of consecutive frames is highly relevant to saliency map detection. Adelson and Bergen [26] proposed an energy model, in which oriented spatio-temporal filters are employed to detect the oriented structures of motion patterns. Along this line, Belardinelli et al. [27] suggested using accumulated spatio-temporal energy to detect saliency maps with the aid of Gabor filtering. Cui et al. [28] presented a temporal generalization of the work by Hou and Zhang [13]. They detected saliency maps by performing *temporal spectral residual* analysis on video slices along $X - T$ and $Y - T$ planes. Tünnermann and Mertsching [29] proposed a region-based approach to saliency detection. They extended region-based attention to the spatiotemporal domain to compile motion saliency, and then presented a biologically inspired system that integrated both spatial and motion saliency to grab static and dynamic stimuli.

The observation lengths of salient regions are nondeterministic. We tackle the foregoing issues by the analysis on *trajectories* obtained from the tracking of feature points in a video, where the trajectories can further be clustered or classified. It is worth noting that trajectories of keypoints have been explored in [21] for anomalous video event detection. In this work, we present a more fast and compact way for trajectory extraction and description. Furthermore, our approach can be used in videos taken by either a stationary or a moving camera.

2) Moving Camera: If the video camcorder is not stationary, the self-movement, or the so-called *ego-motion* induced by the camcorder, should be carefully handled since not all of the motions in videos are salient. Motions induced by the *independent moving objects* (IMOs) [30], [31] are rather crucial. Saliency detection with a moving camera can be achieved by applying the aforementioned techniques for stationary cameras, e.g., [6], [19], [20], [32], if a relatively small ego-motion can be tolerated. Yet this is not always the case, since both strong background motions induced by the camera and foreground motions of IMOs may simultaneously present in videos.

One intuitive way to address this problem is to include ego-motion estimation [31], [33] in preprocessing so that the motions of IMOs can be separated from those of backgrounds. It has been pointed out in [17] that the motions of IMOs tend to be locally consistent. By exploiting this property, the locally consistent motions of IMO as well as the prominent motions of backgrounds can be jointly modeled by structural tensors [34], [35], and the saliency map is subsequently generated. Le Meur et al. [36] assumed that the dominant motion is caused by the movement of cameras and performed hierarchical block matching to estimate the dominant motion. Moreover, the performance of motion estimation may be unstable by various factors, such as illuminations, camera and object motions. Because motion estimation is computationally expensive, Georgiadis et al. [37] skipped motion estimation by directly describing the saliency as the violation of *co-visibility* in terms of epipolar equivalence induced from successive frames. In addition, the trajectory cues can also be applied in a moving camera system like [38] in the sense that trajectories in foregrounds and backgrounds can be further grouped or classified.

All of the above-mentioned frameworks for saliency map detection in video sequences provide solid techniques in task-specific settings of either a stationary or a moving camera. In practice, the camera motions may be continuously varying in some occasions, such as a vehicle event data recorder. We are aware of the explosive growth of vehicle-mounted cameras. There has been a strong demand for saliency detection approaches that can adapt themselves to videos taken by moving and stationary cameras. Compared with previous works, our approach can be distinguished itself by the main feature that it can work with stationary and moving cameras in a unified manner without any prior information regarding camera motions. Therefore, it leads to great flexibility and applicability.

III. Video Saliency Detection

In this section, we describe how to construct saliency maps of a video clip by using extracted trajectories of keypoints. For the sake of clearness, the details of extracting trajectories will be given in the next section. Firstly, a compact and effective trajectory descriptor based on velocity and acceleration entropies is introduced. Then, we present our algorithm that can work with stationary and moving cameras in a unified way and compile the saliency maps by referring to the trajectories.

A. Trajectory Descriptor

The design of our trajectory descriptor builds upon the fact that human beings tend to put attention on moving objects. This fact has been indicated in the seminal work by Egeth and Yantis [16]. Besides, Gao et al. [5] also pointed out that humans also focus on the strong differences between the stimulus at a location and the stimulus in its neighborhood. In videos, the attention is directed to moving objects, which are continuously changing spatial locations. As concluded in [5], one of the representational bases of the visual attention is to separate objects from their locations via motions. The motions of objects can be more effectively represented by continuous trajectories than by the differences of adjacent frames. Cavanagh [39] revealed that low-level signals alone are not responsible when human perceives motions of tracked objects, but the long-range motions attract the attention. Also indicated in [40], when a person can predict the next location of an object, the motion of the object can efficiently guide attention. For a surveillance scenario, objects moving in a straight line take on salience magnitudes that are significantly larger than that of backgrounds [41]. These findings show that human attention focuses on not only early vision but also continuous motions of objects.

With the aim of deriving HVS-consistent saliency maps, our approach simulates HVS by taking the *diversities* of trajectories in both the aspects of velocity and acceleration into account. This is because most important characteristics of motions can be captured by the velocity and acceleration diversities. Specifically, the *entropy*, a measure of the degree of diversities, of each extracted trajectory is considered. These psychological studies [5], [16], [39], [40], [41] support the use of the trajectories in the proposed method.

Assume that K trajectories are extracted in the video, where the kth trajectory T_k consists of a series of keypoints along the space-time volume of the shot. Based on the tracking results, the trajectory T_k can be represented by a time-ordered set of points. Let $\mathbf{x}_k(t) = [u_k(t)\ v_k(t)]^\top$ denote the 2-D coordinate of T_k in frame t. The trajectory T_k is then represented by $\mathbf{X}_k = [\mathbf{x}_k(i)\ \mathbf{x}_k(i+1)\ \ldots\ \mathbf{x}_k(j)]$, where i and j $(j > i)$ are the indices of the first and last frames which T_k resides in, respectively. Please note that trajectories of different lengths, i.e., $(j-i+1)$, are allowed in this representation.

Motivated by the fact that the velocity and acceleration of motions are two perceivable elements in HVS, including these two elements in trajectories generally benefits saliency map detection. The time-ordered representation of T_k supports the efficient computation of velocity $\mathbf{v}_k(t)$ and acceleration $\mathbf{a}_k(t)$ by

$$\mathbf{v}_k(t) = \frac{\mathbf{x}_k(t+1) - \mathbf{x}_k(t)}{\Delta t}, \tag{1}$$

and

$$\mathbf{a}_k(t) = \frac{\mathbf{v}_k(t+1) - \mathbf{v}_k(t)}{\Delta t}, \tag{2}$$

where Δt is the frame time. In this way, a video is represented by a set of trajectories $\{T_k\}$ with $T_k = \{\mathbf{x}_k(t), \mathbf{v}_k(t), \mathbf{a}_k(t)\}$.

In [42], the color probability distribution of each pixel along the temporal axis is introduced. Similar in spirit to [42], we propose the *movement probability distribution* to describe the movement of each trajectory along the temporal domain. In practice, the movement state of each keypoint can vary arbitrarily frame by frame. The diversity of the movement at all keypoints of trajectories represents the motions and the positions of the object where trajectories resides.

To describe the diversity of T_k, the *probability distribution function of the trajectory velocity* is used and defined as follows:

$$p(\mathbf{v}_k(t)) = \frac{||\mathbf{v}_k(t)||}{\sum_{\mathbf{v}_\tau \in T_k} ||\mathbf{v}_k(\tau)|| + \lambda_v}, \tag{3}$$

The *probability distribution function of the trajectory acceleration* can be similarly defined by

$$p(\mathbf{a}_k(t)) = \frac{||\mathbf{a}_k(t)||}{\sum_{\mathbf{a}_\tau \in T_k} ||\mathbf{a}_k(\tau)|| + \lambda_a}, \tag{4}$$

where λ_v and λ_a are small positive constants, and they are empirically set as 10^{-5} for the sake of numerical stability.

With (3) and (4), the *trajectory velocity entropy* and the *trajectory acceleration entropy* of T_k are respectively defined as follows:

$$E_v^{T_k} = - \sum_{\mathbf{v}_k(t) \in T_k} p(\mathbf{v}_k(t)) \log p(\mathbf{v}_k(t)), \tag{5}$$

and

$$E_a^{T_k} = - \sum_{\mathbf{a}_k(t) \in T_k} p(\mathbf{a}_k(t)) \log p(\mathbf{a}_k(t)). \tag{6}$$

$E_v^{T_k}$ in (5) and $E_a^{T_k}$ in (6) respectively represent the diversities of the trajectory T_k in velocity and acceleration along the temporal domain. In our empirical test, motions along the horizontal and vertical axes also give rich information for trajectory analysis. We hence characterize trajectory T_k by a compact (six-dimensional) descriptor $\mathbf{z}_{T_k} = [E_v^{T_k},\ E_{v.x}^{T_k},\ E_{v.y}^{T_k},\ E_a^{T_k},\ E_{a.x}^{T_k},\ E_{a.y}^{T_k}] \in \mathbb{R}^6$, where $E_{v.x}^{T_k}$ is the velocity entropy of T_k along the horizontal axis. $E_{v.y}^{T_k}$, $E_{a.x}^{T_k}$, and $E_{a.y}^{T_k}$ are similarly defined. We will demonstrate in the experiments that such a compact descriptor not only supports fast saliency map generation but also captures sufficient information about the motions in a video. Therefore, it can compile saliency maps of high quality.

B. Trajectory Classification

To utilize the trajectories to compile saliency maps, we aim to pick salient trajectories that correspond to moving objects. According to the study in [16], objects moving at a constant velocity or a constant acceleration attract attention. It implies that the larger the entropy, the more salient the trajectory. However, this property is true only if the camera is fixed. For videos taken by moving cameras, trajectories with large entropies are not necessarily salient. Taking Fig. 5a for an example where a camera focuses on the main character. Thus, the salient parts, the motorbike and the rider, are with weak motions, and thus have lower entropies.

For handling videos recorded by moving and stationary cameras, our approach should be developed upon properties that are invariant to camera motions. It can be observed that trajectories on backgrounds are often consistent, dominant in number, and enclosed by the camera motion no matter whether the camera moves or not. In contrast, the salient trajectories on moving objects in most cases are incompatible with the dominant camera motion. We make use of the properties to separate salient trajectories of moving objects from those of backgrounds. Specifically, we use *one-class SVM* [43], [44] to carry out this idea, and extend the applicability of our approach to handle videos with dominant camera motions. One-class SVM is a classification methodology. It treats positive and negative data in an asymmetrical way. Namely, positive data are similar to each other, while negative data are different in their own ways. In our case, trajectories on backgrounds and on objects are respectively considered the positive and negative data in one-class SVM. In this way, all the trajectories inconsistent with the trajectories under the dominant camera motion are supposed to be classified as negative data. Thus, our approach can be applied to handle both the cases where one or multiple moving objects present in a video.

Suppose that K trajectories, $\{T_i\}_{i=1}^K$, as well as their feature vectors, $\{\mathbf{z}_i\}_{i=1}^K$, are extracted in a video shot. One-class SVM predicts the labels of the trajectories by solving the following constrained optimization problem

$$\begin{aligned} \min_{\mathbf{w}, \{\epsilon_i\}} \quad & \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{C \cdot K} \sum_{i=1}^{N} \epsilon_i - \nu \qquad (7) \\ \text{subject to} \quad & \mathbf{w}^\top \phi(\mathbf{z}_i) \geq \epsilon_i - \nu, \ \text{for } 1 \leq i \leq K, \\ & \epsilon_i \geq 0, \ \text{for } 1 \leq i \leq K, \end{aligned}$$

where C and ν are the two parameters in one-class SVM. We will discuss how to determine their values in the experiments. As a kernel machine, function ϕ maps the data (trajectories) from the input space to some Reproduced Kernel Hilbert Space (RKHS), which is implicitly defined by the adopted kernel. In this work, we select the RBF kernel function for its stable performance, and the inner product of each pair of the mapped data can be efficiently computed via the *kernel trick*. Namely,

$$\begin{aligned} k(\mathbf{z}_i, \mathbf{z}_j) &= \phi(\mathbf{z}_i)^\top \phi(\mathbf{z}_j) & (8) \\ &= \exp\left(\frac{-\|\mathbf{z}_i - \mathbf{z}_j\|^2}{\gamma^2}\right), \ \text{for } 1 \leq i, j \leq K, & (9) \end{aligned}$$

where γ is the hyperparameter. As suggested in [45], we set γ as the mean of the pairwise distances among data $\{\mathbf{z}_i\}_{i=1}^N$.

It turns out that the label of each trajectory is predicted via

$$f(\mathbf{z}_i) = \text{sign}(\mathbf{w}^\top \phi(\mathbf{z}_i) - \nu), \text{ for } 1 \leq i \leq K. \tag{10}$$

LibSVM [46] is adopted in our implementation for both the training and test phases of one-class SVM.

Note that our approach can distinguish itself from previous works that involved the estimation of the dominant camera motions. Since the estimation of the dominant motions is often unreliable and computationally expensive, our approach skips this step, and directly retrieves trajectories incompatible with the dominant motions via one-class SVM. It turns out that the retrieved trajectories reside in moving objects in most cases. Furthermore, our approach makes no assumption about the motions of backgrounds, so it can work with videos with various kinds of camera motions in a unified way.

C. Saliency Map Construction

After completing the learning of one-class SVM, we collect all the trajectories that are predicted as negative to construct the temporal saliency maps. Let $\mathcal{N} = \{T_k\}$ denote the set of the negative trajectories while $[\mathbf{x}_k(t_{k'}) \ \mathbf{x}_k(t_{k'}+1) \ \ldots \ \mathbf{x}_k(t_{k''})]$ denote the coordinates of all the keypoints of $T_k \in \mathcal{N}$ along the temporal axis. We propagate the saliency from trajectories to their surroundings for generating saliency maps. Specifically, for each $T_k \in \mathcal{N}$, the corresponding saliency map S_k is compiled by

$$S_k(\mathbf{x}, t) = \begin{cases} \frac{255}{2\pi\sigma^2} \exp\left(\frac{-\|\mathbf{x}-\mathbf{x}_k(t)\|^2}{2\sigma^2}\right), & \text{if } t_{k'} \leq t \leq t_{k''}, \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

for every time stamp t and every location $\mathbf{x}$ in the video. We repeat the procedure for each trajectory in $\mathcal{N}$. The temporal saliency map $S_\mathcal{T}$ is constructed by

$$S_\mathcal{T}(\mathbf{x}, t) = \min\left(\sum_{T_k \in \mathcal{N}} S_k(\mathbf{x}, t), 255\right). \tag{12}$$

Despite the simplicity in the saliency map construction, as we will demonstrate in experiments, most moving objects, captured by either stationary or moving cameras, are faithfully highlighted in the maps. Besides, the resulting saliency maps are coherent and smooth in adjacent frames. Note that unless further specified, $S_\mathcal{T}$ in (12) is regarded as the saliency maps generated by our approach hereafter.

D. Coupling with Spatial Saliency

The proposed approach explores temporal motions, and identifies salient trajectories by removing alike ones. It can be further improved by integrating with spatial contrast. We suggest that various evidences of saliency are combined in the domain of saliency maps. Specifically, we can impose a frequency-tuned method, such as [7], to compile spatial saliency map $S_\mathcal{S}(\mathbf{x}, t)$ for every location $\mathbf{x}$ and every time stamp t. The fused saliency map $S(\mathbf{x}, t)$ is defined as follows:

$$S(\mathbf{x}, t) = \alpha S_\mathcal{S}(\mathbf{x}, t) + (1-\alpha) S_\mathcal{T}(\mathbf{x}, t), \tag{13}$$

where $\alpha \in [0, 1]$ controls the relative importance between the two kinds of saliency maps.

IV. Trajectory Extraction in A Video

We compute trajectories within each shot, instead of the entire video, to avoid the instability caused by the shot transitions. A shot refers to a sequence of frames, in which object motions in both space and time are continuous. It follows that keypoints which locate on the same object will appear in adjacent frames with similar appearances. It validates the following steps, tracking the keypoints to yield the trajectories of objects. The method in our prior work [47] is used to detect the shots in a video sequence. In the following, we describe how to represent a video by a set of spatially and temporally coherent trajectories. It consists of the three stages, including 1) keypoint detection, 2) keypoint description, and 3) keypoint tracking. Each stage is detailed in the following.

A. Keypoint Detection

The goal of this stage is to efficiently and stably detect keypoints in a given video. One can apply any off-the-shelf corner detectors, e.g., [48], [49], to each frame of the video to extract keypoints. Among various corner detection algorithms, the *FAST* detector [49] has shown its superior results in both accuracy and speed. It implements a heuristic rule and a machine learning concept so that the non-corner points can be efficiently filtered out.

However, the keypoints obtained via frame by frame detection may be unstable due to various factors, such as the continuous changes of video content and the motions of the camera. These unfavorable factors often cause false detection and missed detection between adjacent frames. This problem is known as the *repeatability problem* [49] for corner detection. During processing videos, the state-of-the-art detector, FAST, still suffers from this problem. This is a serious issue in our case, since unstable keypoint detection results in the breaks of trajectories. To solve the problem, we propose a *spatial-temporal FAST* (ST-FAST) detector, which combines the temporal information between adjacent frames with the spatial information to detect keypoints.

For a 2-D image point $\mathbf{x}$, FAST considers $\mathbf{x}$ as a corner by examining neighbor pixels on a circle surrounding $\mathbf{x}$. Let $I(\mathbf{x})$ denote the intensity of $\mathbf{x}$. If the intensities of most of the neighbor pixels are brighter than $I(\mathbf{x}) + \theta$, $\mathbf{x}$ is then a corner. θ is a positive offset to avoid the inference of noise. If the intensities of most of the neighbor pixels are darker than $I(\mathbf{x}) - \theta$, $\mathbf{x}$ is also a corner.

Comparing the intensity of $\mathbf{x}$ with those of its neighbors provides the spatial information to determine if $\mathbf{x}$ is a corner or not. In this paper, we further explore the temporal information to identify if a point $\mathbf{x}(t)$, which is located at $\mathbf{x}$ in the tth frame, is a keypoint. We treat the video as a 3-D spatio-temporal volume with each $\mathbf{x}(t)$ as a 3-D voxel. For each voxel $\mathbf{x}(t)$, we consider the set of its neighboring voxels as the union of the voxels that locate on a $r \times r$ tube centered on $\mathbf{x}(t)$. To examine if $\mathbf{x}(t)$ is a spatio-temporal extreme or not, one of the three states is assigned to each $\mathbf{v} \in \mathcal{N}_{\mathbf{x}(t)}$ by

$$S(\mathbf{v}) = \begin{cases} d, & \text{if } I(\mathbf{v}) < I(\mathbf{x}(t)) - \theta, \\ b, & \text{if } I(\mathbf{v}) > I(\mathbf{x}(t)) + \theta, \\ s, & \text{otherwise,} \end{cases} \tag{14}$$

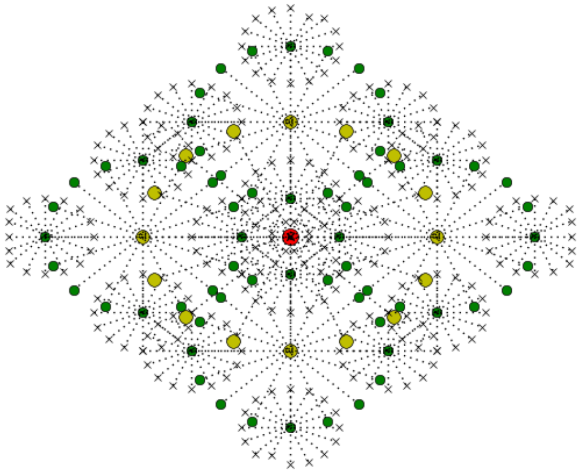


Fig. 1. The galaxy sampling pattern [50].

where $I(\mathbf{v})$ and $I(\mathbf{x}(t))$ are the intensities of $\mathbf{v}$ and $\mathbf{x}(t)$, respectively. If most of the voxels in $\mathcal{N}_{\mathbf{x}(t)}$ have state d or have state b, $\mathbf{x}(t)$ is regarded as a detected keypoint, i.e., a spatio-temporal FAST corner. Since the temporal evidences have been taken into account in verifying keypoints, the repeatability problem in adjacent frames is significantly alleviated.

B. Keypoint Description

After retrieving keypoints, the next step is to match keypoints between adjacent frames for generating long-term trajectories. Researches on invariant local descriptors, e.g., [48], [50], [51], are quite extensive. These descriptors can effectively solve the matching problem between images. Based on the property of repeatability in the detected keypoints, these descriptors are suitable for matching and tracking the keypoints of objects in consecutive video frames. Specifically, we use a new binary descriptor *galaxy* [50] to match keypoints between adjacent frames because it is robust to lighting changes and has good performance. The descriptor consists of a galaxy sampling pattern and a binary encoding scheme.

The structure of the galaxy descriptor contains four levels including the center of the galaxy (level 0), the fixed stars of the galaxy (level 1), the planets of the solar systems (level 2) and the satellites of the planets (level 3). As shown in Fig. 1, a keypoint $\mathbf{x}(t)$ represents the center G (red circle) of the galaxy. Several fixed stars F (yellow circle) surround G. Each fixed star may contain several planets P (green circle) to form its own solar system. Satellites S (black cross) may exist for each planet.

The galaxy sampling pattern is applied to describing the local region R centered on each detected keypoint $\mathbf{x}(t)$. According to the hierarchical structure in the pattern, we firstly set the location of G as $\mathbf{x}(t)$, and then sequentially determine the sampled pixels in F, P, and S. The galaxy sampling pattern defines which pixels in R will be sampled to construct the descriptor. Let $\{\mathbf{x}_{i,l}(t)\}$ be the set of the sampled pixels, where l and i denote the level and the index of the pixels, respectively. Each sampled pixel $\mathbf{x}_{i,l}(t)$ will be compared to $\mathbf{x}_{j,l}(t)$ that has the same ancestor and is on its opposite side. We call $(\mathbf{x}_{i,l}(t), \mathbf{x}_{j,l}(t))$ an *opposite pair* hereafter.

As indicated in [52], color information often enhances the discriminability of descriptors. Thus, the galaxy descriptor is computed on the *transformed color space* $\{R', B', G'\}$ [52], which has been shown to be an efficient and effective color invariant space. A naïve way is to apply the descriptor to each color channel independently, and encode the relative relationships in each of the opposite pairs. Benefiting from that the transformed color space supports cross-channel comparisons, we can further increase the discriminative power of the descriptor by retrieving information across color channels. Specifically, the cross-channel relative relationship between an opposite pair $(\mathbf{x}_{i,l}(t), \mathbf{x}_{j,l}(t))$ is defined as

$$d(C_m(\mathbf{x}_{i,l}(t)), C_n(\mathbf{x}_{j,l}(t))) = \begin{cases} 1, & C_m(\mathbf{x}_{i,l}(t)) < C_n(\mathbf{x}_{j,l}(t)), \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

where $C_m \in \{R', G', B'\}$ and $C_n \in \{R', G', B'\}$. $C_m(\mathbf{x}_{i,l}(t))$ is the pixel value at the location $\mathbf{x}_{i,l}(t)$ in channel C_m. Similarly, $C_n(\mathbf{x}_{j,l}(t))$ is defined.

For a specific keypoint $\mathbf{x}(t)$ as well as two channels C_m and C_n, we now define the feature vector $D_{C_m,C_n}(\mathbf{x}(t))$ generated by the galaxy descriptor. Let $\mathcal{U}$ denote the set of the opposite pairs in all the three levels F, P, and S. Feature vector $D_{C_m,C_n}(\mathbf{x}(t))$ corresponds to a binary string, whose value is computed as follows:

$$D_{C_m,C_n}(\mathbf{x}(t)) = \sum_{i=1}^{|\mathcal{U}|} 2^{i-1} \times d(C_m(\mathbf{x}_{i_1,i_l}(t)), C_n(\mathbf{x}_{i_2,i_l}(t))), \tag{16}$$

where the ith opposite pair in $\mathcal{U}$ consists of two points $\mathbf{x}_{i_1,i_l}(t)$ and $\mathbf{x}_{i_2,i_l}(t)$ in level i_l. $|\mathcal{U}|$ is the cardinality of $\mathcal{U}$.

Since each of C_m and C_n belongs to $\{R', G', B'\}$, there are totally nine cross-channel combinations. By excluding the redundant ones, we consider six channel combinations, including $\{(R', R'), (R', G'), (R', B'), (G', G'), (G', B'), (B', B')\}$. The final feature vector $D(\mathbf{x}(t))$ extracted by the galaxy descriptor is the concatenation of the six binary strings, each of which corresponds to one particular cross-channel combination.

C. Keypoint Tracking

Since the galaxy descriptor characterizes detected keypoints in form of binary strings, *Hamming distance* can be used to very efficiently measure the similarity between keypoints. For compiling trajectories in a given video, we firstly extract the ST-FAST keypoints for each shot. For a keypoint in frame $t-1$, we find the most similar keypoint in frame t via Hamming distance. If the spatial distance between the two keypoints is small enough, we deem that the corresponding point in frame t has been found for the keypoint in frame $t-1$. Based on the matching results, we can track each keypoint of the video. The linked list of each tracked keypoint yields a trajectory. By repeating the procedure, the trajectories in the video are extracted.

TABLE I
DATASET SUMMARY.

Index	Video	Frame #	Resolution	Camera	Object #
1	Motorbike	79	320 × 240	Moving	1
2	Auto_race	21	320 × 240	Moving	1
3	Soccer	74	320 × 240	Moving	5
4	Surveillance	106	320 × 240	Stationary	4
5	Crossroad	3019	320 × 240	Stationary	51
6	Indoor	564	320 × 240	Rotation+Scale	1

V. EXPERIMENTAL RESULTS

A comprehensive study of the performance evaluation and analysis for the proposed approach is conducted in this section.

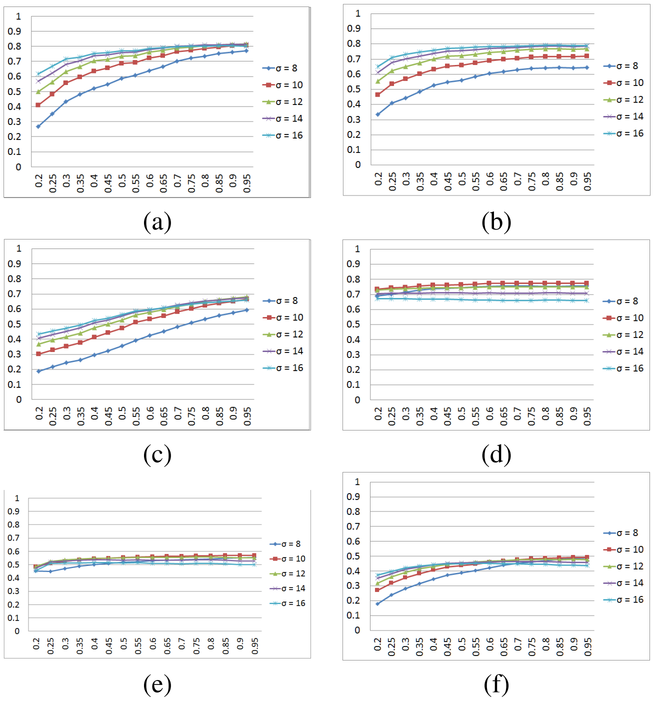
A. Dataset

We collected and manually annotated six video sequences for evaluating approaches to video saliency detection. The first three and the last videos were captured by moving cameras, so dominant camera motions presented in the four videos. The fourth and the fifth videos were taken by stationary cameras. The first video `motorbike` captured a motorbike rider on the road. The fixation point focuses on the rider. Therefore, the rider as well as the motorbike keeps stationary, while the backgrounds, such as the sky and the bus, move. The second video is a clip of auto racing. Similar to the first video, the camera in the video `auto_race` focused on a white car, and tracked it. The first two videos contain one single salient object. In contrast, the third video `soccer` captured several salient objects, where the soccer players are walking in the soccer field. The focus of the camera followed them. It is worth pointing out that the motions of the soccer players are inconsistent, and their observation lengths are also different. The fourth video `surveillance` was selected from the PETS2001 dataset [53], and it was captured by a stationary surveillance camera. A white car moved along the road and several people walked from the bottom-left side to the right-hand side. As a surveillance video, moving foregrounds are considered as salient objects. To evaluate the proposed method in long-term surveillance videos with illumination changes, we used the `crossroad` dataset [54]. The last video `indoor` included camera rotation and scale changes of the foreground object at the same time. The foreground object moved toward the camera and thus caused scale changes. In addition, the object turned right and the camera rotated to keep on the focus of the object. It is a very challenge dataset because both the rotation of the camera and scale change of the object involve in the video clip.

The properties of these videos are summarized in TABLE I. Some frames and their annotated ground truth of these videos are shown in Fig. 5 ∼ Fig. 10, respectively. These videos span a wide spectrum of practical conditions, such as different camera settings and motions, numbers of salient objects, and the observation lengths. Thus, the dataset can serve as a good test bed for performance evaluation and analysis.

B. Evaluation Metrics

We use the same evaluation metrics suggested in [55] to measure the quality of a pixel-level saliency map. Let $n\text{TP}$ denote the number of true positives, which are the detected salient pixels. Let $n\text{FP}$ and $n\text{FN}$ represent the numbers of false positives and false negatives, respectively. The false positives here correspond to the non-salient pixels that are predicted salient, while the false negatives correspond to the salient pixels that are predicted non-salient. With $n\text{TP}$, $n\text{FP}$ and $n\text{FN}$, the *precision* and the *recall* of a saliency map can be computed. While the former is the fraction of the detected pixels that are salient, the latter is the fraction of the salient pixels that are detected. Specifically, we have



(a) (b)
(c) (d)
(e) (f)

Fig. 2. The F-measure scores (y-axis) by our approach with different combinations of parameter ν (x-axis) in OCSVM and parameter σ in saliency propagation. The six figures correspond to videos (a) `motorbike`, (b) `auto_race`, (c) `soccer`, (d) `surveillance`, (e) `crossroad`, and (f) `indoor`, respectively.

$$\text{PRECISION} = \frac{n\text{TP}}{n\text{TP} + n\text{FP}}, \tag{17}$$

and

$$\text{RECALL} = \frac{n\text{TP}}{n\text{TP} + n\text{FN}}. \tag{18}$$

Precision and recall are two conflicting goals. One can trivially optimize one of them by ignoring the other. Thus, precision and recall are usually referred jointly for performance evaluation. For the sake of compactness, we use *F-measure* to consider precision and recall at the same time. Its definition is given as follows:

$$\text{F-MEASURE} = 2 \cdot \frac{\text{PRECISION} \cdot \text{RECALL}}{\text{PRECISION} + \text{RECALL}}. \tag{19}$$

We will use F-measure as the primary metric in the experiments, though precision and recall are also given for reference.

C. Parameter Selection

In our approach, two main parameters will affect the results of temporal saliency map generation. The first one is ν in

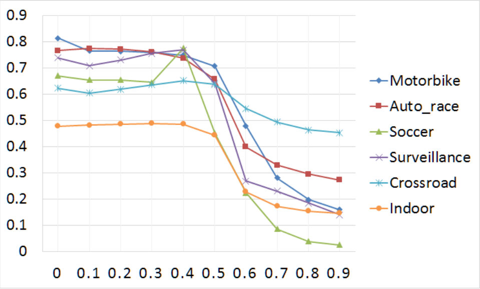


Fig. 3. The F-measure scores (y-axis) of the combined saliency maps with respect to α (x-axis) in six videos.

one-class SVM (OCSVM) in (7), which controls the upper bound on the fraction of training errors. The other one is the band width σ of the Gaussian mask in (11) when propagating saliency from a trajectory to its neighborhood. The OCSVM tends to predict trajectories negative (salient here) with larger values of ν. Thus, larger ν will result in higher recall and lower precision. On the other hand, σ determines the extents of saliency propagation. Larger σ means that the saliency will be diffused to a broader neighborhood. Therefore, larger σ also leads to higher recall as well as lower precision.

We compiled the saliency maps by applying our approach with various combinations of the two parameters to the six videos. It is difficult to find out the optimal ranges of the parameters by merely referring to either precision or recall. We hence adopt F-measure, which is the harmonic mean of the recall and precision rates. As shown in Fig. 2, the values of F-measure converge toward the optimal results, if ν is larger than 0.85. Moreover, the higher F-measure values are achieved on all the six videos, when $\sigma = 12$ and $\nu \geq 0.8$. For a fair evaluation, we set $\nu = 0.9$ and $\sigma = 12$, and report the results in all the following experiments.

To evaluate the effect of α in (13) on combining the temporal and spatial saliency maps, we varied the value of α from 0 to 0.9. Fig. 3 shows the F-measure scores of the combined saliency maps with respect to different values of α. Note that when α equals to zero, the combined saliency maps are the same as the temporal saliency maps. As shown in Fig. 3, combining the temporal and spatial saliency maps achieve worse F-measure scores in videos `motorbike` and `auto_race`, because the spatial contrast of backgrounds is visually significant in colors and edges. In contrast, human beings can focus on foreground objects based on the temporal cues. Thus, how to properly combine temporal and spatial saliency maps remains a problem. Here, we set $\alpha = 0.4$ in the experiments. It indicates that we put more emphasis on temporal cues. This finding is consistent to [56], which also suggests higher weights over temporal cues.

D. Comparative Baselines

The primary comparison of interest is to validate whether our approach is comparable to the state-of-the-art approaches to video saliency detection. In addition, we would like to identify the contributions of each developed component in this work. To this end, our approach is compared with the state-of-the-art methods including the image-based frequency-tuned (FT) approach [7] and two video-based methods using phase spectrum of quaternion Fourier transform (PQFT) [24] and textural contrast (TC) [56]. We also compared our methods with two famous background subtraction methods: the first one is based on Gaussian mixture model (GMM) [18] and the second one is Vibe [57]. In addition, some variants of our approach are considered to identify the contributions of some individual components in our approach. We respectively replaced the proposed ST-FAST keypoint detector by other spatial-temporal keypoint detectors, including space-time interest points (STIP) [58] (denoted by **Ours w STIP**) and FAST [49] (**Ours w FAST**). We also replaced one-class SVM (OCSVM) by a degenerate variant **Ours w/o OCSVM**, in which thresholding is used to classify the entropy of a trajectory. We computed the average entropy of each trajectory in velocity and acceleration. Trajectories with higher entropies are considered salient. Another variant, which fuses both the temporal and spatial saliency maps, is referred as **Ours w SC**.

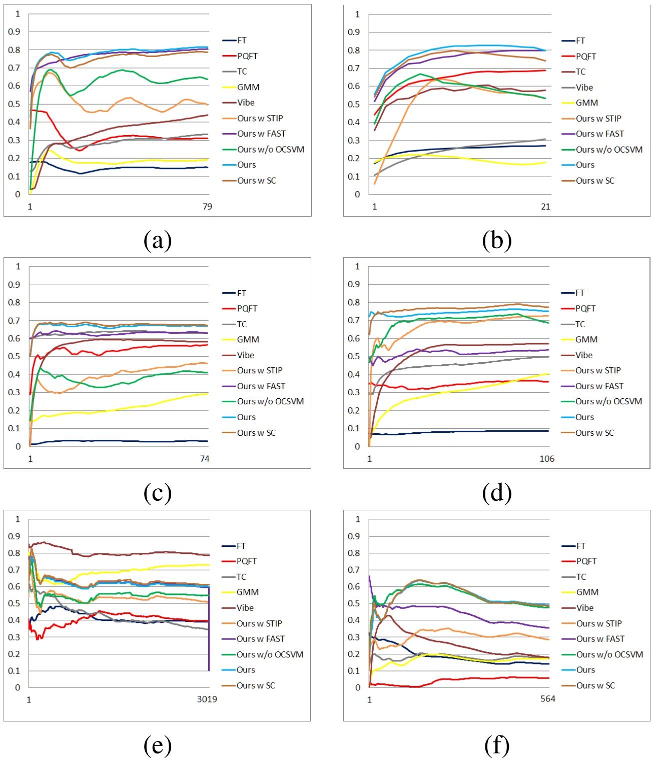
E. Quantitative Results

TABLE II shows the recall, precision and F-measure scores by applying the baselines and our approach to the video clips. Note that we compute the nTP, nFP and nFN of all of the frames at first, and then compute the recall, precision and F-measure scores. For each dataset, our method (**Ours**) or its variant (**Ours w SC**) achieves the highest F-measure scores except video `crossroad`. The F-measure scores of FT [7], PQFT [24] and TC [56] methods are lower due to the lack of long-term motions. GMM [18] and Vibe [57] are background modeling methods which assume that the cameras are static. Thus, when the cameras move as shown in videos `motorbike`, `auto_race`, `soccer` and `indoor`, their F-measure scores are significantly lower than those of the proposed method. Nevertheless, Vibe [57] achieved the best F-measure scores in video `crossroad`. Although baselines **Ours w STIP** and **Ours w FAST** also take long-term motions into consideration, the instability of the STIP and FAST keypoints leads to lower F-measure scores. Thus, it is important to extract stable spatial-temporal keypoints for video saliency map generation. Also shown in TABLE II, the F-measure scores of baseline **Ours w/o OCSVM** are lower in the videos with dominant camera motions, because it simply applies thresholding to the entropy of a trajectory, and hence cannot account for dominant camera motions. Our method (**Ours**) and the variant (**Ours w SC**) perform well in most cases.

Fig. 4 shows the frame-wise F-measure scores for each video, respectively. As we expect, the results of FT [7] are worse than those of the other methods. Such results indicate the importance of motion information for video saliency map generation. In contrast, the motion based methods [24] and [56] get better results by imposing the motion information between two frames. Nevertheless, the long-term motion information, such as trajectories, is not considered. As a result, its performance is worse than that of the proposed

TABLE II
PERFORMANCES, IN PRECISION, RECALL, AND F-MEASURE, OF THE BASELINES AND OUR APPROACH.

Video	Evaluation metric	FT	PQFT	TC	GMM	Vibe	Ours w STIP	Ours w FAST	Ours w/o OCSVM	Ours	Ours w SC
Motorbike	Precision	0.102	0.851	0.190	0.308	0.381	0.617	0.702	0.932	0.814	0.787
	Recall	0.243	0.439	0.538	0.134	0.333	0.472	0.952	0.478	0.812	0.727
	F-Measure	0.144	0.579	0.808	0.281	0.180	0.340	0.522	0.632	**0.813**	0.748
Auto_race	Precision	0.139	0.871	0.131	0.203	0.211	0.722	0.642	0.935	0.852	0.831
	Recall	0.210	0.583	0.923	0.185	0.498	0.401	0.928	0.342	0.695	0.667
	F-Measure	0.167	0.699	0.228	0.193	0.294	0.507	0.759	0.501	**0.766**	0.737
Soccer	Precision	0.110	0.780	0.513	0.933	0.686	0.763	0.562	0.916	0.719	0.857
	Recall	0.014	0.568	0.910	0.124	0.495	0.271	0.638	0.264	0.626	0.462
	F-Measure	0.026	0.657	0.656	0.218	0.568	0.394	0.598	0.410	0.670	**0.775**
Surveillance	Precision	0.695	0.286	0.390	0.968	0.590	0.618	0.306	0.887	0.637	0.599
	Recall	0.047	0.486	0.592	0.180	0.475	0.742	0.994	0.555	0.878	0.775
	F-Measure	0.088	0.360	0.468	0.301	0.518	0.672	0.468	0.683	0.738	**0.768**
Crossroad	Precision	0.943	0.517	0.221	0.784	0.654	0.577	0.735	0.766	0.718	0.861
	Recall	0.267	0.342	0.945	0.692	0.917	0.500	0.544	0.433	0.550	0.525
	F-Measure	0.415	0.404	0.354	0.735	**0.763**	0.533	0.624	0.552	0.622	0.651
Indoor	Precision	0.154	0.080	0.106	0.158	0.111	0.171	0.236	0.368	0.369	0.417
	Recall	0.128	0.045	0.371	0.184	0.457	0.831	0.707	0.704	0.664	0.579
	F-Measure	0.140	0.057	0.165	0.170	0.178	0.284	0.354	0.483	0.477	**0.485**



(a) (b)

(c) (d)

(e) (f)

Fig. 4. The F-Measure scores of the comparative baselines and the proposed method along the frames of videos (a) `motorbike`, (b) `auto_race`, (c) `soccer`, (d) `surveillance`, (e) `crossroad`, and (f) `indoor`.

TABLE III
NUMBERS OF BACKGROUND AND FOREGROUND TRAJECTORIES.

Video	Background	Foreground	Total
Motorbike	1412	156	1568
Auto_race	618	68	686
Soccer	1550	170	1720
Surveillance	6440	1176	7616
Crossroad	16048	1783	17831
Indoor	3316	368	3684

TABLE IV
AVERAGE COMPUTATION TIME PER FRAME.

Video	TE (s)	SD (s)	Total (s)
Motorbike	0.0338	0.0018	0.0356
Auto_race	0.0398	0.0018	0.0416
Soccer	0.0337	0.0016	0.0353
Surveillance	0.0347	0.0013	0.0360
Crossroad	0.0578	0.0097	0.0675
Indoor	0.0548	0.0035	0.0583

method. Because cameras have dominant motions in video clips `motorbike`, `auto_race`, `soccer` and `indoor`, the F-measure scores of GMM [18] and Vibe [57] are lower than those of our method. In contrast, GMM and Vibe achieve better performance in clips `surveillance` and `crossroad`, since these cameras are static, and the constraints of background modeling are satisfied.

As shown in Fig. 4d, the F-measure scores of **Ours w FAST** are much worse than those of the proposed method, because **Ours w FAST** only extracts keypoints frame by frame, and neglects temporal consistency. Thus, the jittering of keypoint locations between adjacent frames easily occurs, and corrupts the computation of the entropies. As a result, false alarms occur due to the instability of the keypoints. Similar problems occur in **Ours w STIP**. Because **Ours w/o OCSVM** applies a simple threshold for retrieving trajectories with higher entropies as salient trajectories, false alarms frequently occur in baseline **Ours w/o OCSVM** for moving cameras. The proposed method considers the long-term motion information, so it can extract more accurate video saliency maps. More importantly, our method (**Ours** with the aid of OCSVM is applicable to videos taken by either a moving or a stationary camera without any prior information.

TABLE III reports the numbers of the extracted foreground (salient) and background trajectories. There are more background trajectories. OCSVM in these cases can effectively remove the dominant and consistent background trajectories.

The proposed method is implemented by using Microsoft Visual C++ 2010 and OpenCV 2.4.4 on a personal computer with an Intel Core i7 3.4GHz CPU and 16GB main memory. We did not perform any additional hardware optimization or support, such as GPU, in the implementation. TABLE IV shows the average running time per frame of the proposed method with respect to trajectory extraction (TE) and one-class SVM based video saliency map detection (SD). The proposed method supports real-time detection of saliency maps, because it effectively applies the binary descriptors [50] for trajectory extraction. Moreover, the low-dimensional (six-dimensional) representation of each trajectory not only faithfully models the motions of trajectories but also significantly reduces the running time of one-class SVM. Because more trajectories are extracted in `crossroad`, more computation time is required. Nevertheless, our method detects about $15 \sim 29$ frames per second.

F. Detected Saliency Maps

To gain insight into the quantitative results, some detected saliency maps of the six videos are displayed in Fig. 5 $\sim$ Fig. 10, respectively. In Fig. 5, the first two columns show the frame indices and the corresponding frames in video `motorbike`. The detected ST-FAST keypoints are marked in Fig. 5c. These keypoints were repeatedly detected along the video frames. This property facilitates the generation of spatially and temporally coherent trajectories. The manually labeled ground truth is shown in Fig. 5d, in which regions in white denote saliency.

Fig. 5e, Fig. 5f, and Fig. 5g give the saliency maps generated by FT [7], PQFT [24], and TC [56], respectively. Although FT can effectively discover salient regions in still images, it fails in this video. This is because the important information of motions is ignored. The complex backgrounds of high contrast, such as the sky and the road, caused the numerous false positives. In contrast, the motorbike and the rider can be detected by PQFT, because motion features extracted in adjacent frames were taken into account. The temporal information in TC is computed from the differences of the current frame and the next three frames, so the lamppost of the background in the 58th frame is detected. The detected salient objects of both PQFT and TC are broken into pieces due to the lack of long-term temporal information.

As we expect, the saliency maps shown in Fig. 5h and Fig. 5i, which are detected by the two background modeling methods GMM [18] and Vibe [57], are broken. Because the camera moves and tracks the rider and the motorcycle, the backgrounds change with time. As a result, the two methods fail to model the backgrounds that are assumed to be static, and obtain broken saliency maps.

The saliency maps yielded by the variants of our approach, **Ours w STIP** and **Ours w FAST**, are displayed in Fig. 5j and Fig. 5k, respectively. STIP [58] considers only the gradients of keypoints in the x, y and t directions. In contrast, ST-FAST takes the neighbor voxels surrounding a keypoint into account, so it can then detect more reliable keypoints. Because the FAST detector identifies keypoints frame by frame, the locations of the same keypoints of an object in adjacent frames may jitter. Although we can still track the keypoints using STIP and FAST to generate trajectories, such instability will affect the computation of entropies and result in sub-optimal saliency detection results. Fig. 5l shows the results by another variant, **Ours w/o OCSVM**, in which trajectories are classified by thresholding their velocity and acceleration entropies. Implied by [16], the entropies in our trajectory descriptor help the detection of objects that move with a constant velocity or acceleration, but this property holds only in videos taken by stationary cameras. Thus, the results given in Fig. 5l are not good enough. In addition, the background clutters may be moving at a more constant speed than the foreground objects. Nevertheless, they are not identified as salient regions. This is because background trajectories in the videos are of short lengths, and the foreground trajectories are of relatively longer lengths. According to the definitions of (5) and (6), trajectories of longer lengths tend to have larger entropies. As a result, background trajectories can be separated by baseline **Ours w/o OCSVM**.

The results of our method are shown in Fig. 5m. Compared to the baselines, our method can more precisely identify both the rider and the motorcycle as salient objects. It can be observed in this video that most of the extracted trajectories in the backgrounds are consistently enclosed by the dominant camera motion. Moreover, the trajectories in the backgrounds are more than those in the foregrounds. We leveraged the two observations, and employed one-class SVM to remove the background trajectories. It turns out that the remaining trajectories faithfully reveal the salient regions, i.e., the rider and the motorbike. Fig. 5n shows the results of combination of the proposed temporal saliency map and spatial saliency map detected by [7]. Because the spatial saliency map contains many non-salient backgrounds, the combined results are not as good as the results of the temporal saliency map. This founding indicates that it needs to find a good way to properly combine temporal and spatial cues instead of simply combining both cues by addition as shown in [36], [56].

Fig. 6 shows some of the detected saliency maps in video `auto_race` where the camera followed a racing car. Similar to the results in Fig. 5e, the race field was also detected as saliency by FT owing to the strong color contrast in it. Both the methods PQFT and TC took the short-term motion features to yield better outcomes. However, the salient object is still broken as shown in Fig. 6f and Fig. 6g. GMM and Vibe again failed to detect the salient race car as shown in Fig. 6h and Fig. 6i, respectively. As shown in Fig. 6l, baseline **Ours w/o OCSVM** still suffered from the same problem, i.e., the camera motion, and resulted in the unsatisfactory saliency maps. Our approach and its variant **Ours w FAST** performed well in the video. The white car was accurately detected in the saliency maps as shown in Fig. 6m and Fig. 6k. It is worth noting that a corner recognized by the ST-FAST detector must be recognized by the FAST detector, but not vice verse. Thus, the extracted trajectories in baseline **Ours w FAST** are usually denser. It follows that there are more false positives in **Ours w FAST**, while there are more false negatives in our approach. The railing in the 5th and 8th frames and the black windshield in the 14th frame are the examples.

Detecting saliency maps in video `soccer` is challenging in the sense that there are multiple moving soccer players, each of which has a distinct pattern of motion. Even the motions of the local parts of a player are different. Fig. 7

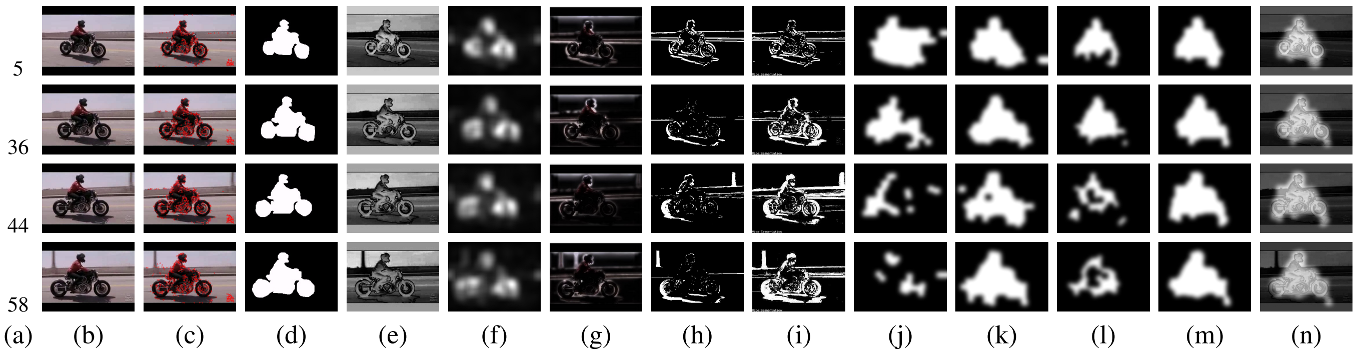


Fig. 5. A few saliency maps of video `Motorbike`. (a) Frame indices. (b) Videos frames. (c) Detected ST-FAST keypoints. (d) Annotated ground truth. (e) ∼ (i) The saliency maps detected by various approaches, including (e) FT [7], (f) PQFT [24], (g) TC [56], (h) GMM [18], (i) Vibe [57], (j) **Ours w STIP**, (k) **Ours w FAST**, (l) **Ours w/o OCSVM**, (m) **Ours**, and (n) **Ours w SC**.

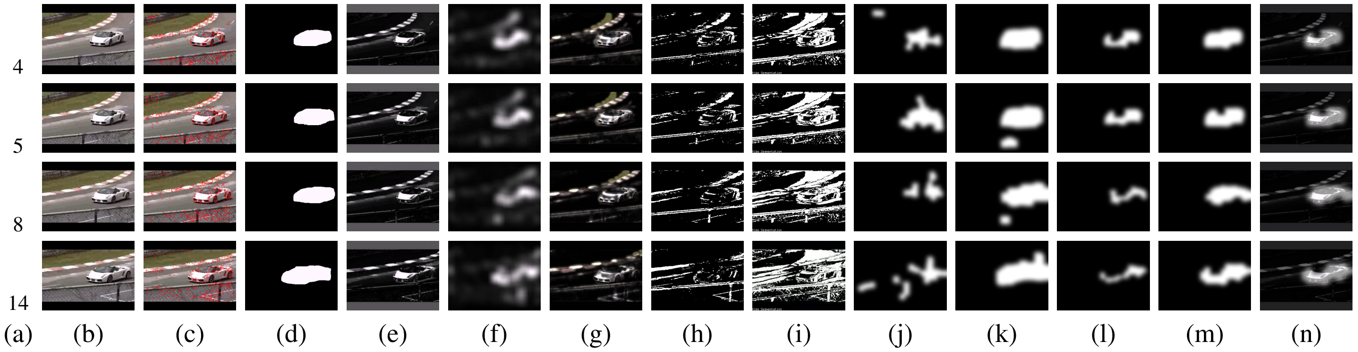


Fig. 6. A few saliency maps of video `Auto_race`. (a) Frame indices. (b) Videos frames. (c) Detected ST-FAST keypoints. (d) Annotated ground truth. (e) ∼ (i) The saliency maps detected by various approaches, including (e) FT [7], (f) PQFT [24], (g) TC [56], (h) GMM [18], (i) Vibe [57], (j) **Ours w STIP**, (k) **Ours w FAST**, (l) **Ours w/o OCSVM**, (m) **Ours**, and (n) **Ours w SC**.

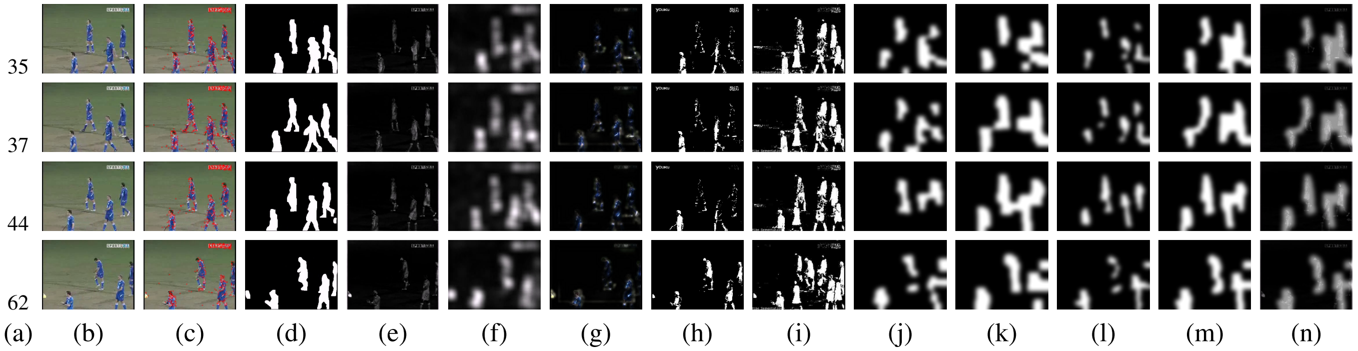


Fig. 7. A few saliency maps of video `soccer`. (a) Frame indices. (b) Videos frames. (c) Detected ST-FAST keypoints. (d) Annotated ground truth. (e) ∼ (i) The saliency maps detected by various approaches, including (e) FT [7], (f) PQFT [24], (g) TC [56], (h) GMM [18], (i) Vibe [57], (j) **Ours w STIP**, (k) **Ours w FAST**, (l) **Ours w/o OCSVM**, (m) **Ours**, and (n) **Ours w SC**.

shows some of the saliency maps compiled by the baselines and our approach, respectively. It can be observed in Fig. 7 that the baselines caused either numerous false positives or numerous false negatives. The saliency maps in Fig. 7m by our approach and Fig. 7n by **Ours w SC** are still HVS-consistent. This is because our approach is designed in a general way. It makes no assumption about the motion patterns of moving objects. Since the backgrounds in this video are enclosed by the dominant camera motion, one-class SVM in our approach can still identify the trajectories in the backgrounds. In addition, the developed ST-FAST detector is helpful in extracting trajectories in the regions with complex motion, since it suppresses noises in each single frame by checking temporal consistence across frames.

The fourth and the fifth videos are surveillance videos selected from the PETS2001 dataset [53] and [54], respectively. As shown in Fig. 8e and Fig. 9e, the image-based method [7] did not accurately detect moving vehicles and pedestrians due to the lack of motion information, though it successfully suppressed false negatives in the backgrounds. The video-based methods PQFT, TC, and our variant **Ours w STIP** and **Ours w FAST** suffered from the problem of instability caused by the change of lighting as well as the low quality of the video. There were many false positives in the resulting saliency

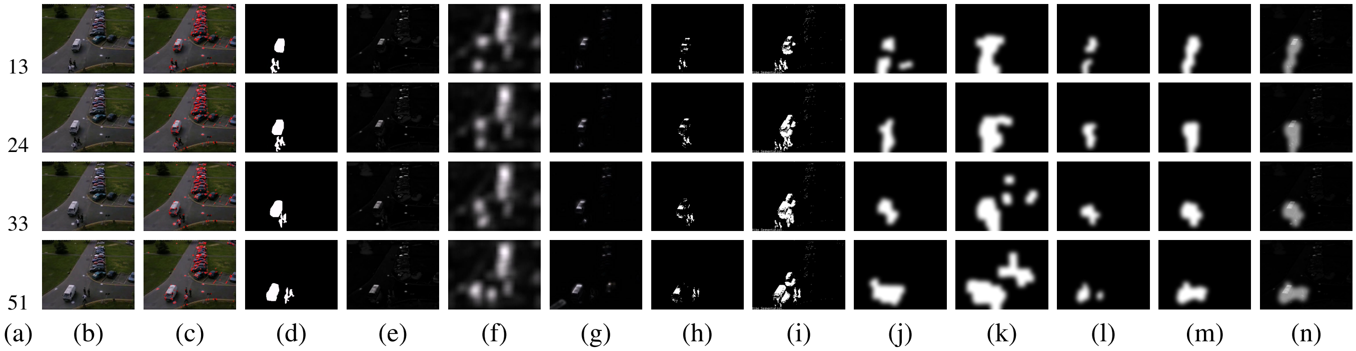


Fig. 8. A few saliency maps of video `surveillance`. (a) Frame indices. (b) Videos frames. (c) Detected ST-FAST keypoints. (d) Annotated ground truth. (e) ∼ (i) The saliency maps detected by various approaches, including (e) FT [7], (f) PQFT [24], (g) TC [56], (h) GMM [18], (i) Vibe [57], (j) **Ours w STIP**, (k) **Ours w FAST**, (l) **Ours w/o OCSVM**, (m) **Ours**, and (n) **Ours w SC**.

maps, which can be seen in Fig. 8f and also in Fig. 9f and 9g. Since GMM and Vibe are background modeling methods, their performance is expected to outperform methods based on saliency detection in surveillance videos. Compared with the results of GMM, Vibe achieved better foreground extraction results as shown in Fig. 8i and Fig. 9i. Our approach explored long-term object motions that were captured by tracking ST-FAST corners. The yielded saliency maps, given in Fig. 8m and Fig. 9m, are much better. The good performance of **Ours w/o OCSVM** in videos taken by stationary cameras manifests the effectiveness of the proposed compact trajectory descriptor.

Fig. 10 shows the detected saliency maps of the sixth video, `indoor`. Because the person walked from the door to the camera and turned right, significant scale variation of the person presented in the video. It leads to the difficulty of consistently identifying the person as a salient foreground object. Moreover, when the person turned, the camera also rotated to track the person at the same time. It resulted in two different camera motions in the video. As a results, all of the compared methods failed to identify the person as a salient object. In contrast, **Ours** and **Ours w SC** can still identify the person as shown in Fig. 10m and Fig. 10n, respectively. Note that the motions of background trajectories are still consistent and dominant, when cameras zoom or rotate. In addition, the motions of foreground trajectories are different from those of background ones. Thus, our approach still works in the cases, since it removes the consistent and dominant trajectories by one-class SVM, and uses the remaining trajectories to compile the saliency map.

The promising results on these testing video sequences demonstrate the effectiveness and large applicability of our approach. It is also worth noting that our approach does not require any pre-training and any prior knowledge about videos. Thus, our approach also provides an efficient solution to background modeling. Our demo video is available at http://www.cs.nchu.edu.tw/∼crhuang/file/VSD.avi.

VI. Conclusions

We have presented an effective approach that utilizes long-term trajectory activities to construct the video saliency maps, and is applicable to videos taken by stationary and moving cameras. In this approach, features invariant to camera motions are exploited by one-class SVM, and are used to identify the salient trajectories that are in most cases incompatible with the dominant camera motion. We have also introduced a compact trajectory descriptor based on motion diversity, and it allows us to detect moving objects of various observation lengths. Besides, the ST-FAST detector, which locates corners by jointly using spatial and temporal cues, is designed to stabilize the tracking of keypoint trajectories. Our approach elegantly combines these components, and leads to spatially and temporally coherent saliency maps.

In the future, we will use the yielded saliency maps, and investigate their impacts on video content analysis, such as video abstraction, retrieval and event recognition. Recently, we are also aware the explosive growth of vehicle-mounted cameras. Videos captured by vehicle-mounted cameras often contain continuously varying camera motions. That is, the dominant camera motion is smoothly changing when the vehicle is moving, and is stationary when the vehicle stops. We would like to leverage the merits of our approach in efficiency, effectiveness, and great applicability, and apply it to analyzing videos taken by vehicle-mounted cameras.

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson, "Top-down control of visual attention in object detection," in *Proc. Int'l Conf. Image Processing*, 2003.

[3] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems*, 2006.

[4] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in Neural Information Processing Systems*, 2008.

[5] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, 2008.

[6] V. Mahadevan and N. Vasconcelos, "Background subtraction in highly dynamic scenes," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2008.

[7] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[8] C. Kanan, M. Tong, L. Zhang, and G. Cottrell, "SUN: Top-down saliency using natural statistics," *Visual Cognition*, vol. 17, no. 6-7, pp. 979–1003, 2009.

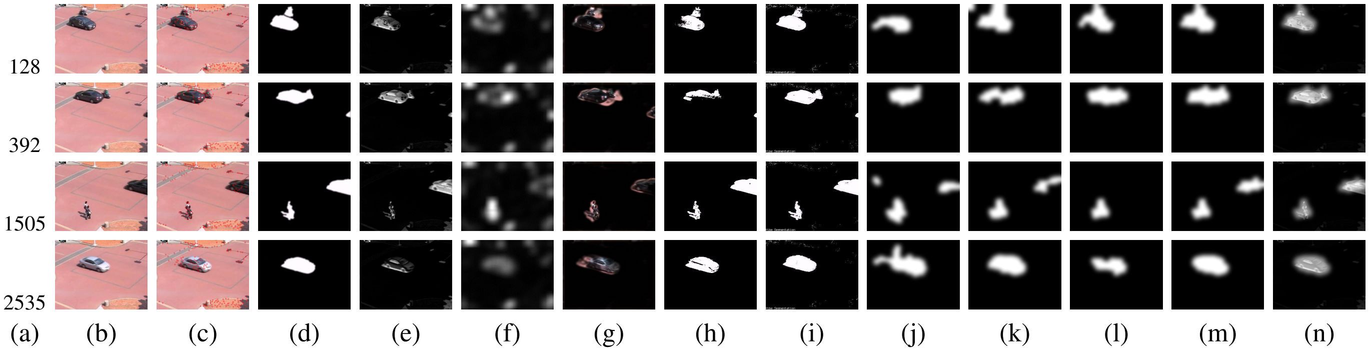


Fig. 9. A few saliency maps of video `crossroad`. (a) Frame indices. (b) Videos frames. (c) Detected ST-FAST keypoints. (d) Annotated ground truth. (e) ∼ (i) The saliency maps detected by various approaches, including (e) FT [7], (f) PQFT [24], (g) TC [56], (h) GMM [18], (i) Vibe [57], (j) **Ours w STIP**, (k) **Ours w FAST**, (l) **Ours w/o OCSVM**, (m) **Ours**, and (n) **Ours w SC**.

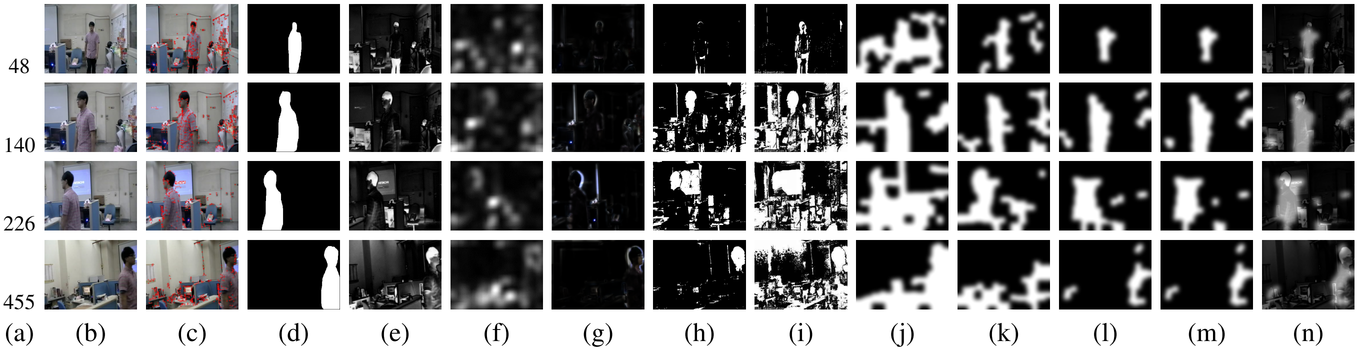


Fig. 10. A few saliency maps of video `indoor`. (a) Frame indices. (b) Videos frames. (c) Detected ST-FAST keypoints. (d) Annotated ground truth. (e) ∼ (i) The saliency maps detected by various approaches, including (e) FT [7], (f) PQFT [24], (g) TC [56], (h) GMM [18], (i) Vibe [57], (j) **Ours w STIP**, (k) **Ours w FAST**, (l) **Ours w/o OCSVM**, (m) **Ours**, and (n) **Ours w SC**.

[9] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011.

[10] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[11] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 989–1005, 2009.

[12] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[13] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2007.

[14] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[15] S. Lu and J.-H. Lim, "Saliency modeling from image histograms," in *Proc. European Conf. Computer Vision*, 2012.

[16] H. E. Egeth and S. Yantis, "Visual attention: Control, representation, and time course," *Annual Review of Psychology*, vol. 48, pp. 269–297, 1997.

[17] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[18] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. Conf. Computer Vision and Pattern Recognition*, 1999.

[19] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Proc. European Conf. Computer Vision*, 2000.

[20] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, "Background modeling and subtraction of dynamic scenes," in *Proc. Int'l Conf. Computer Vision*, 2003.

[21] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1544–1554, 2008.

[22] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2008.

[23] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 446–456, 2011.

[24] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[25] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Conf. Multimedia*, 2006, pp. 815–824.

[26] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *Journal of Optical Society of America, A*, vol. 2, no. 2, pp. 284–299, 1985.

[27] A. Belardinelli, F. Pirri, and A. Carbone, "Motion saliency maps from spatiotemporal filtering," in *IEEE Workshop Applications of Computer Vision*, 2009, pp. 112–123.

[28] X. Cui, Q. Liu, and D. N. Metaxas, "Temporal spectral residual: Fast motion saliency detection," in *Proc. ACM Conf. Multimedia*, 2009, pp. 617–620.

[29] J. Tünnermann and B. Mertsching, "Region-based artificial visual attention in space and time," *Cognitive Computation*, pp. 1–19, 2013.

[30] A. Strehl and J. K. Aggarwal, "Detecting moving objects in airborne forward looking infra-red sequences," in *Proc. IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications*, 1999.

[31] Y. Zhang, S. J. Kiselewich, W. A. Bauson, and R. Hammoud, "Robust moving object detection at distance in the visible spectrum and beyond using a moving camera," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2006.

[32] Y. Sheikh, O. Javed, and T. Kanade, "Background subtraction for freely moving cameras," in *Proc. Int'l Conf. Computer Vision*, 2009.

[33] A. Jazayeri, H. Cai, J.-Y. Zheng, and M. Tuceryan, "Vehicle detection and tracking in car video based on motion model," *IEEE Trans. Intelligent Transportation Systems*, vol. 12, no. 2, pp. 583–595, 2011.

[34] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, 1981.

[35] S. Li and M.-C. Lee, "An efficient spatiotemporal attention model and its application to shot matching," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 17, no. 10, pp. 1383–1387, 2007.

[36] O. Le Meur, D. Thoreau, P. Le Callet, and D. Barba, "A spatio-temporal model of the selective human visual attention," in *Proc. Int'l Conf. Image Processing*, 2005, vol. 3, pp. 1188–1191.

[37] G. Georgiadis, A. Ayvaci, and S. Soatto, "Actionable saliency detection: Independent motion detection without independent motion estimation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[38] S. Wu, O. Oreifej, and M. Shah, "Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories," in *Proc. Int'l Conf. Computer Vision*, 2011.

[39] P. Cavanagh, "Attention-based motion perception," *Science*, vol. 257, no. 5076, pp. 1563–1565, 1992.

[40] A. P. Hillstrom and S. Yantis, "Visual motion and attentional capture," *Perception and Psychophysics*, vol. 55, no. 4, pp. 399–411, 1994.

[41] L. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," vol. 22, no. 8, pp. 774–780, 2000.

[42] Y. F. Ma and H. J. Zhang, "Detecting motion object by spatio-temporal entropy," in *Proc. Int'l Conf. Multimedia and Expo*, 2001.

[43] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[44] L. Manevitz and M. Yousef, "One-class SVMs for document classification," *J. Machine Learning Research*, pp. 139–154, 2002.

[45] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, 2002.

[46] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[47] C.-R. Huang, H.-P. Lee, and C.-S. Chen, "Shot change detection via local keypoint matching," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 1097–1108, 2008.

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

[49] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 105–119, 2010.

[50] G.-H. Huang and C.-R. Huang, "Binary invariant cross color descriptor using galaxy sampling," in *Proc. Int'l Conf. Pattern Recognition*, 2012, pp. 2610–2613.

[51] C.-R. Huang, C.-S. Chen, and P.-C. Chung, "Contrast context histogram - an efficient discriminating local descriptor for object recognition and image matching," *Pattern Recognition*, vol. 41, no. 10, pp. 3071–3077, 2008.

[52] K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[53] PETS'01 Database, "ftp://ftp.pets.rdg.ac.uk/pub/PETS2001," 2001.

[54] C.-R. Huang, H.-C. Chen, and P.-C. Chung, "Online surveillance video synopsis," in *Proc. Int'l Symposium on Circuits and Systems*, 2012, pp. 1843–1846.

[55] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1937–1944.

[56] W. Kim and C. Kim, "Spatiotemporal saliency detection using textural contrast and its applications," *To appear in , IEEE Trans. on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2013.

[57] O. Barnich and M.-V. Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[58] I. Laptev, "On space-time interest points," *Int'l J. Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

Chun-Rong Huang received the B.S. and Ph.D. degrees in the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, in 1999 and 2005, respectively.

In 2005, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, where, since 2005, he has been a Postdoctoral Fellow. In 2010, he became an Assistant Professor with both the Institute of Networking and Multimedia, and the Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan.

His research interests include computer vision, computer graphics, multimedia signal processing, image processing, and medical image processing. Dr. Huang is a member of the IEEE Circuits and Systems Society and the Phi Tau Phi honor society.

Yun-Jung Chang received the B.S. degree in the Department of Computer Science and Information Engineering, National Chi Nan University, Taiwan, in 2012. He is currently pursuing his master's degree in the Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan. He is also a research assistant of Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan since 2012.

Zhi-Xiang Yang received the B.S. degree in the Department of Computer Science, National Chiao Tung University, Taiwan, in 2010, and the M.S. degree in the Institute of Networking and Multimedia, National Chung Hsing University, Taichung, Taiwan, in 2012. He was a research assistant at the Research Center for Information Technology Innovation, Academia Sinica, Taiwan from 2010 to 2012.

Yen-Yu Lin received the B.S. degree in information management in 2001, the M.S. and Ph.D. degrees in computer science and information engineering in 2003 and 2010 respectively, all from National Taiwan University. He is currently an assistant research fellow at the Research Center for Information Technology Innovation, Academia Sinica, Taiwan. His research interests include computer vision, pattern recognition, and machine learning.