SCIENCE CHINA
Information Sciences

April 2020, Vol. 63 140307:1–140307:15
https://doi.org/10.1007/s11432-019-2784-4

• RESEARCH PAPER •
Special Focus on Deep Learning in Remote Sensing Image Processing

Deep feature extraction and motion representation for satellite video scene classification

Yanfeng GU[1], Huan LIU[1], Tengfei WANG[1], Shengyang LI[2,3] & Guoming GAO[1*]

[1]*School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin* 150001, *China;*
[2]*Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing* 100094, *China;*
[3]*The Key Laboratory of Space Utilization, Chinese Academy of Sciences, Beijing* 100094, *China*

Received 1 November 2019/Revised 8 January 2020/Accepted 9 February 2020/Published online 9 March 2020

Abstract Satellite video scene classification (SVSC) is an advanced topic in the remote sensing field, which refers to determine the video scene categories from satellite videos. SVSC is an important and fundamental step for satellite video analysis and understanding, which provides priors for the presence of objects and dynamic events. In this paper, a two-stage framework is proposed to extract spatial features and motion features for SVSC. More specifically, the first stage is designed to extract spatial features for satellite videos. Representative frames are firstly selected based on the blur detection and spatial activity of satellite videos. Then the fine-tuned visual geometry group network (VGG-Net) is transferred to extract spatial features based on spatial content. The second stage is designed to build motion representation for satellite videos. The motion representation of moving targets in satellite videos is first built by the second temporal principal component of principal component analysis (PCA). Second, features from the first fully connected layer of VGG-Net are used as high-level spatial representation for moving targets. Third, a small network of long and short term memory (LSTM) is further designed for encoding temporal information. Two-stage features respectively characterize spatial and temporal patterns of satellite scenes, which are finally fused for SVSC. A satellite video dataset is built for video scene classification, including 7209 video segments and covering 8 scene categories. These satellite videos are from Jilin-1 satellites and Urthecast. The experimental results show the efficiency of our proposed framework for SVSC.

Keywords satellite videos, classification, convolutional neural network, CNN, long and short term memory, LSTM, motion representation

Citation Gu Y F, Liu H, Wang T F, et al. Deep feature extraction and motion representation for satellite video scene classification. Sci China Inf Sci, 2020, 63(4): 140307, https://doi.org/10.1007/s11432-019-2784-4

1 Introduction

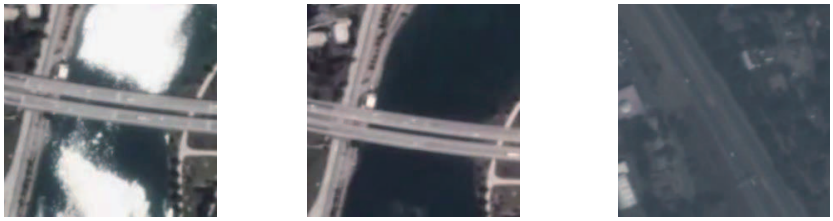
Recently, the launch of very high-resolution video satellites has enabled us to observe moving targets on the Earth surface, which can provide ultra high-definition (UHD) videos with about 1 m spatial resolution. Compared with the traditional remote sensing static images, satellite videos can achieve real-time dynamic observations for a certain area, which have a wide range of applications in disaster monitoring, traffic monitoring, suspicious object surveillance, and ocean resource monitoring.

Satellite video scene classification (SVSC) is a fundamental challenge in the goal of automated image and video understanding. The ability to distinguish scenes is very useful as it can provide priors for the presence of objects [1–3] or dynamic events [4]. For example, detect moving vehicles in satellite videos, where SVSC can first provide a detection area in a large scale video, such as kinds of roads and parking lots. In satellite videos, dynamically abnormal events usually happen in specific scene areas, e.g., gas

* Corresponding author (email: ggm@hit.edu.cn)

(a) (b) (c) (d)

Figure 1 (Color online) Various occlusion phenomena in satellite videos. (a) Moving targets are occluded by the shadow of high buildings in a parking lot scene; (b) a highway scene is occluded by dense trees; (c) moving targets are occluded when they pass through the bottom of the overpass; (d) a highway scene is occluded by thin clouds.



(a) (b) (c)

Figure 2 (Color online) Low imaging quality in satellite videos. (a) Overexposure in a bridge scene; (b) a reference frame of (a) without overexposure; (c) low contrast in a highway scene.

plume pollution in industrial areas. In addition, SVSC is also valuable for the management of video browsing, retrieval and caption [5]. Moreover, SVSC can distinguish different motion patterns for urban traffic dynamic analysis, which provides a large scale tool for monitoring and understanding typical events and actions compared with local camera videos in computer science. This paper will concentrate on the issue of SVSC in the remote sensing field.

Critical challenges to SVSC arise from occlusion phenomena and the limits of imaging quality, which cause larger diversity for satellite videos compared with general videos. Figure 1 shows sample frames from the dataset introduced in this paper that highlights such diversity caused by various occlusion phenomena. For instance, owing to parallax effect between high buildings and satellites, high buildings occlude moving targets as shown in Figure 1(a). Dense plants and overpasses partially or totally occlude moving targets as shown in Figure 1(b) and (c). On the other hand, drifting thin clouds cover certain scenes as shown in Figure 1(d). The limits of imaging quality mainly include overexposure and low contrast as shown in Figure 2(a) and (c), which bring larger intra-class difference.

Scene classification from static image has been intensively studied. Hand-crafted feature descriptors have been widely applied for this task, such as scale-invariant feature transform (SIFT), local binary pattern (LBP), and histogram of oriented gradients (HOG). These basic feature descriptors can be further coded by bag of words (Bow) [6], fisher vector (FV) [7] and sparse coding [8]. Owing to lack of labeled training samples, knowledge transferring is successfully applied for scene classification [9], e.g., transferring pre-trained convolutional neural network (CNN) in the remote sensing filed. Pre-trained CNNs can be used as feature descriptors to extract hierarchical features [10–12], which are dominant methods for aerial scene classification. To fully utilize features from different layers of pre-trained CNNs, various fusion strategies are proposed to enhance classification performance [13–15].

Compared with image scene classification, there are motion information besides appearance information for video scene classification. Considering appearance features, basic feature descriptors are consistent with image features, e.g., Bow [16]. To utilize dynamic information, image-based features can be captured in spatiotemporal orientation instead of extracting in spatial orientation, such as LBPs from three orthogonal planes (LBP-TOP) [17], 3D SIFT [18], and spatiotemporal oriented energy (SOE) [19]. Some popular methods to describe motion information are based on optical flow, such as motion boundary histograms (MBH) and histogram of optical flow (HOF) [20]. Methods based on dense trajectories

(DT) [21] and improved DT (iDT) [22] have been viewed as the standard of hand-crafted features for video scene classification. Recently, deep learning has been applied for video scene classification which focuses on how to deal with the temporal dimension. The first group extends 2D CNN to 3D CNN with 3D convolution and 3D pooling, which captures discriminative features along both spatial and temporal dimensions while maintaining a certain temporal structure [23–25]. Tran et al. [26] proposed convolutional 3D (C3D) features based on 3D CNN, which achieved good performance for dynamic scene classification. The second group extracts motion features, e.g., 2D dense optical flow maps, embedded into CNN networks [27]. The third group combines CNN with a temporal sequence modeling, such as recurrent neural network (RNN) [28], long short-term memory (LSTM) [29, 30], and bidirectional RNN (B-RNN) [31]. Feichtenhofer et al. [32] proposed a temporal residual network (T-ResNet), which is fully convolution in spatiotemporal orientation by temporal residual units. For different kinds of deep learning models, performance of a method can be boosted by information fusion from multiple cues and models.

In this paper, a two-stage framework is proposed to extract spatial features and temporal features for SVSC. The first stage is to extract spatial features. With regard to most of stable and unchanged observation areas during one lasting observation time, representative frames are firstly selected based on the blur detection and the spatial activity of satellite videos. Then the fine-tuned visual geometry group network (VGG-Net) [33] is transferred to extract spatial features for satellite videos. It is similar with feature extraction for image scenes. The second stage is to extract temporal features for satellite videos. Then motion representation of moving targets is extracted by the second temporal principal component of principal component analysis (PCA). Then the VGG-Net combined with LSTM further encodes motion information in the time dimensionality. Finally, spatial features and motion representative features are fused for SVSC. In addition, it is necessary to analyze moving targets in satellite videos. Motion representative features are used for SVSC considering moving targets, where the criteria is whether moving targets exist in satellite videos. The two-stage framework is verified on the built satellite video scene dataset.

The remainder of this paper is organized as follows: Section 2 describes the proposed framework for satellite video classification. First, selection of representative frames for SVSC and motion representation of moving targets are introduced. Then, it presents the transferring framework of VGG-Net for feature extraction and classification. And the basic theory of LSTM is provided. Section 3 introduces the dataset for satellite video classification and presents experimental results based on the proposed framework. Finally, the conclusion is drawn in Section 4.

2 Methods description

Our classification task is to classify different satellite video scenes. The proposed framework for SVSC is shown in Figure 3. The classification task is divided into two main stages. The first stage is to extract spatial features based on scene content. The second stage is to build motion representation. In the first stage, one single frame is firstly selected as the representative frame for a satellite video scene. And the fine-tuned VGG-Net is transferred as the basic architecture to extract spatial features from full-connected layers. In the second stage, the main steps include: (1) motion information of consecutive frames is extracted by temporal principal components for satellite videos; (2) each principal component frame extracted in (1) is respectively as inputs of VGG-Net to extract high-level spatial information; (3) then temporal features are encoded by stacked two layers of LSTM. Finally, spatial features and temporal features are stacked for feature fusion, and a softmax layer is used as a classifier for SVSC.

2.1 Selection of representative video frames

In processing video tasks, generally, not all frames are processed because of huge computation. In terms of temporal redundancy, sampling along with temporal dimensionality is a general processing strategy [28, 29]. In addition, representative frames are also extracted as abstracted representation of video content for video indexing, browsing, and retrieval [34, 35]. For SVSC, most scenes are basically

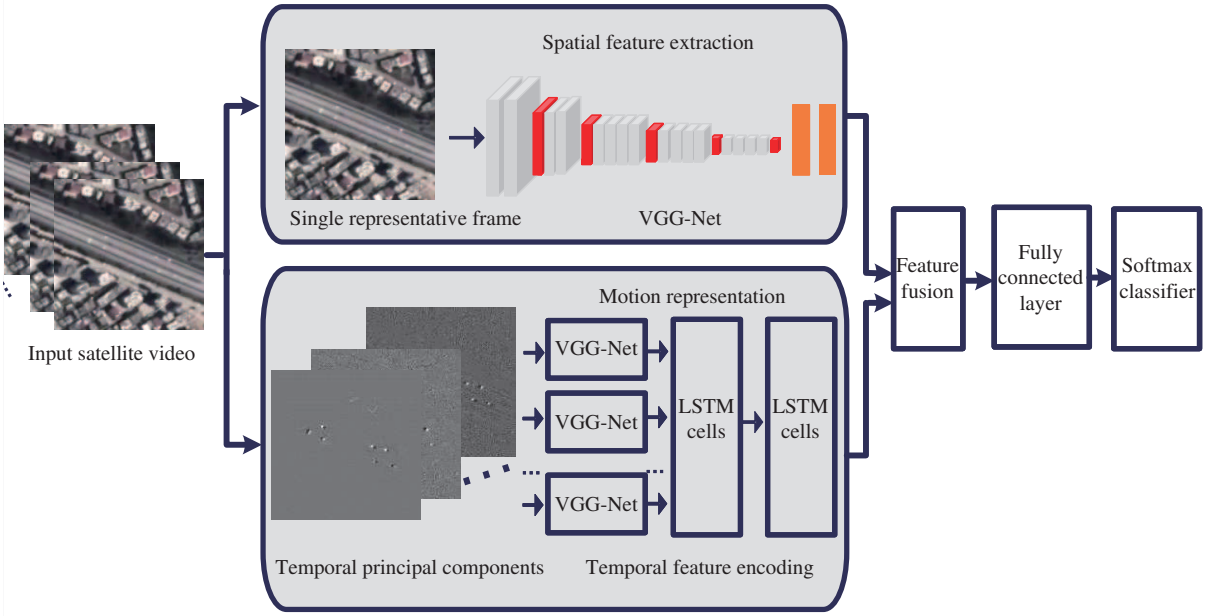


Figure 3 (Color online) The architecture of the proposed framework for satellite video scene classification (SVSC). Spatial features are extracted by the VGG-Net based on representative frames. Motion representation is built based on temporal principal components, followed by VGG-Net and LSTM for temporal feature encoding. Then spatial features and temporal features are fused. Finally, a softmax classifier is used for SVSC.

unchanged owing to large-scale Earth observation. In this case, some frames as still images can best represent content of satellite video for scene classification while reducing data volume. Base on this, only color features are used for selecting representative frames without taking motion cues of moving objects into consideration [36]. In addition, the phenomenon of blur appears frequent in satellite videos, which impacts quality of satellite videos. Therefore, both the blur detection and spatial activity of videos are considered in this paper to select representative frames.

Let $\boldsymbol{I} : \left(\Omega \subset \mathbb{R}^2\right)$ be a frame of size of $m \times n$ pixels. Moreover, $\boldsymbol{x} := (x, y)^{\mathrm{T}}$ denotes a pixel in the image domain Ω. The blur metric of frame $\boldsymbol{I}$ can be computed according to [37], where blur metric is measured by comparing neighboring pixels of a blurred image $\boldsymbol{B}$ with $\boldsymbol{I}$. In terms of variation between $\boldsymbol{B}$ and $\boldsymbol{I}$, a low value means the frame $\boldsymbol{I}$ is already blurred whereas a high value means the frame $\boldsymbol{I}$ is sharp. Let $\boldsymbol{H}_v$ and $\boldsymbol{H}_h$ respectively denote a horizontal and vertical motion filter to create $\boldsymbol{B}_v$ and $\boldsymbol{B}_h$ to model the blur effect.

$$\boldsymbol{B}_v = \boldsymbol{I} * \boldsymbol{H}_v, \quad \boldsymbol{B}_h = \boldsymbol{I} * \boldsymbol{H}_h. \tag{1}$$

where $*$ denotes convolution operator.

The absolute difference images $\boldsymbol{D\_I}_v$ and $\boldsymbol{D\_I}_h$ are computed by

$$\boldsymbol{D\_I}_v = \mathrm{abs}\left(\boldsymbol{I}\left(i, j\right) - \boldsymbol{I}\left(i-1, j\right)\right), \tag{2}$$

$$\boldsymbol{D\_I}_h = \mathrm{abs}\left(\boldsymbol{I}\left(i, j\right) - \boldsymbol{I}\left(i, j-1\right)\right), \tag{3}$$

where $i = 1, \ldots, m-1$ and $j = 1, \ldots, n-1$. Then the variations of neighboring pixels after blur are evaluated by

$$\boldsymbol{V}_v = \max\left(0, \boldsymbol{D\_I}_v - \mathrm{abs}\left(\boldsymbol{B}_v\left(i, j\right) - \boldsymbol{B}_v\left(i-1, j\right)\right)\right), \tag{4}$$

$$\boldsymbol{V}_h = \max\left(0, \boldsymbol{D\_I}_h - \mathrm{abs}\left(\boldsymbol{B}_h\left(i, j\right) - \boldsymbol{B}_h\left(i, j-1\right)\right)\right). \tag{5}$$

Finally, the blur value is measured by comparing the sum variations of neighboring pixels after blur with the original frame. The normalized result is as follows:

$$\mathrm{Blur} = \max\left(1 - \frac{\sum \boldsymbol{V}_v}{\sum \boldsymbol{D\_I}_v}, 1 - \frac{\sum \boldsymbol{V}_h}{\sum \boldsymbol{D\_I}_h}\right), \tag{6}$$

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

Figure 4 (Color online) Temporal principal components of consecutive video frames. (a) and (b) a pair of consecutive video frames for a highway satellite video scene; (c) moving targets in frame (a) are outlined with red circles; (d) the first temporal principal component; (e) the second temporal principal component.

where the range of blur value is from 0 to 1, which are respectively the best and the worst quality in terms of blur effect.

Then the entropy of each frame is used to measure spatial activity. Let i denote the grayscale value and $p\left(i,k\right)$ be the probability of i in frame k.

$$H\left(k\right)=-\sum_{i=1}^{n}p\left(i,k\right)\log_2\left(p\left(i,k\right)\right). \tag{7}$$

The basic principle of selecting representative frames is to ensure a low blur metric (to avoid motion blur) and a high spatial activity. These two measures are combined to compute for each frame as

$$S\left(k\right)=w_H\frac{H\left(k\right)}{\sigma_H}-w_B\frac{\mathrm{Blur}\left(k\right)}{\sigma_B}, \tag{8}$$

where σ_H and σ_B are respectively the standard deviation of the entropy H and blur metric Blur. And w_H and w_B are weights of two measures. The frame with the highest value is selected as the representative frame for later classification.

2.2 Motion information extraction for moving targets in satellite videos

The second stage is building motion representation for satellite videos. Therefore, how to describe moving information is crucial for this task. The significant motion is mainly caused by moving targets, which corresponds to various transportation. In satellite videos, non-motion related patterns of significance, such as flicker (caused by noise or illumination intensity change), waves, and adjustments from overexposure to normal states, are interfering factors for motion representation. To overcome these interfering factors, PCA is applied for extracting significant motion in satellite videos, which represents video content in low temporal dimensionality [38] and reduces interference from unrelated motion.

Let $N\times T$ matrix $\boldsymbol{X}$ be the input, containing the spatial features along with time dimensionality (T is duration in videos and $N\gg T$). We implement PCA through singular value decomposition (SVD) as follows:

$$\boldsymbol{X}=\boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}}, \tag{9}$$

where both $\boldsymbol{V}$ and $\boldsymbol{U}$ are orthonormal matrixes, and Σ is a diagonal matrix whose diagonal elements are non-negative singular values sorted in a descending order. Projecting the video matrix $\boldsymbol{X}$ into each column of $\boldsymbol{V}$ can yield the corresponding temporal principal components (TPCs).

Moving targets in satellite videos are very small and background scenes are basically unchanged. For consecutive frames, this kind characteristic of moving targets and background can be distinguished by different TPCs. The top TPCs represent background scene in the consecutive frames, the subsequent temporal principal components mainly contain moving targets, and remaining TPCs with small eigenvalues mainly capture noise in videos.

Motion representation of consecutive video frames are shown in Figure 4, where PCA is implemented to two consecutive video frames. Different dominant information has been represented by different TPCs of consecutive video frames. The first TPC is the main background scene information. And the second TPC

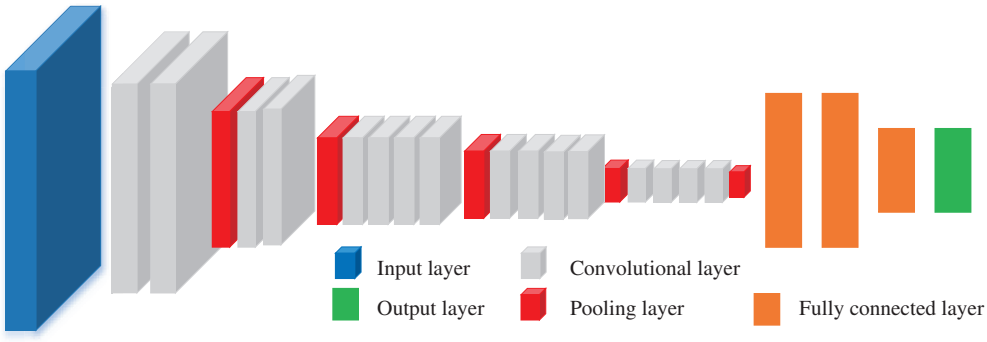


Figure 5 (Color online) The spatial feature extraction network based on the VGG-Net architecture, including 5 blocks of convolutional layers followed by pooling layers and 3 fully connected layers.

can denote the moving targets with some noise. There are 12 moving targets in the pair of consecutive video frames. Motion displacements of 7 moving targets have been correctly extracted, where one moving object has blocked another object, leading to an integral motion representation. The remaining moving targets are missed owing to low contrast density between vehicles and roads. These problems, such as multiple small moving targets, occlusion problem, lack of rich texture information, and noise interference, bring challenges to extract motion information. In the part of extracting motion information for moving targets, the second TPC (Se-TPC) of two consecutive video frames is used to represent moving targets in satellite videos.

2.3 Spatial feature extraction based on the deep VGG-Net

Representative spatial features are crucial for SVSC. In terms of the superiority of CNN to hand-craft features, classical architectures of CNN trained on large dataset in computer science, such as AlexNet, CaffeNet and VGG-Net, have been successfully transferred for scene classification of very high resolution (VHR) images in the remote sensing field [10, 12, 39]. Similar with VHR images, spatial scene content in satellite videos can be also extracted by transferred CNN architectures. In our proposed framework, VGG-Net pre-trained on the ImageNet dataset is transferred as the base network architecture for spatial feature extraction. In the first stage for spatial feature extraction based on representative frames, the weights of pre-trained VGG-Net are used as the initial weights, and then fine-tuned weights based on the satellite video data set. In the second stage for motion representation, the pre-trained weights of VGG-Net are fixed as a high-level feature extractor, where outputs of the second fully connected layer as spatial feature representation for moving targets in Se-TPC. The VGG-Net used in this paper has 19 weight layers, which comprises five blocks of convolutional layers and three fully connected layers, where each block is followed by one pooling layer [33]. In this architecture, the input size of images is resized to $224 \times 224 \times 3$, and the size of the last fully connected layer is modified with 8 nodes based on our dataset instead of 1000 nodes for ImageNet classification task, where nodes represent number of classes to be classified. The architecture of VGG-Net used in this paper is shown in Figure 5.

2.4 Temporal feature coding and classification

To further capture global temporal dependency of motion representation for SVSC, LSTM is applied for temporal feature coding. LSTM neural network has received increasing attention as a general sequence processing mechanism for video classification [29, 30]. The key of LSTM is memory units that allow LSTM to learn when to forget previous hidden states and when to update hidden states given new information. This mechanism of LSTM can learn long-term dependencies without suffering from vanishing and exploding gradients as traditional RNN [29].

The architecture of memory unit is shown in Figure 6. In our framework, $\boldsymbol{x}_t$ denotes feature representation of moving targets in satellite videos, which is encoded by VGG-Net based on the Se-TPC as shown in Figure 3. The aim of LSTM is to map input sequence $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ through hidden vector

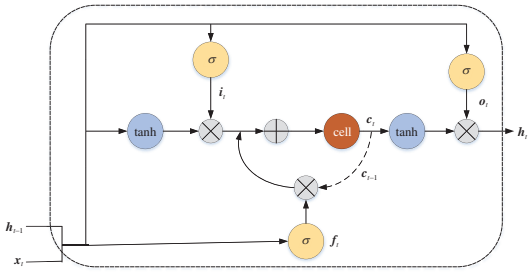


Figure 6 (Color online) A diagram of an LSTM memory unit.

sequence $(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T)$ to output vector sequence $(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T)$ by iterating the following equations from $t = 1$ to T:

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ci}\boldsymbol{c}_{t-1} + \boldsymbol{b}_i\right), \tag{10}$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cf}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f\right), \tag{11}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tanh\left(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c\right), \tag{12}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{c}_t + \boldsymbol{b}_o\right), \tag{13}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh\left(\boldsymbol{c}_t\right), \tag{14}$$

where $\boldsymbol{i}, \boldsymbol{f}, \boldsymbol{o}$ and $\boldsymbol{c}$ are respectively the input gate, forget gate, output gate, and cell active vectors. σ is the logistic sigmoid function, and $\odot$ is an element-wise product operator. The $\boldsymbol{W}$ terms denote weight matrices (e.g., $\boldsymbol{W}_{hi}$ is the hidden-input weight bias), and the $\boldsymbol{b}$ terms denote bias vectors (e.g., $\boldsymbol{b}_i$ is the input bias vector). At each time step, the LSTM unit receives inputs from two external sources. One is the feature $\boldsymbol{x}_t$ from current frame. The other source is the previous hidden states of all LSTM units in the same layer $\boldsymbol{h}_{t-1}$. In addition, each gate has an internal source, e.g., the cell state $\boldsymbol{c}_{t-1}$. Because $\boldsymbol{i}_t$ and $\boldsymbol{f}_t$ are modulated by sigmoid function, their values lie within the range [0,1], and LSTM unit learns to selectively forget its previous memory or consider its current input according to $\boldsymbol{i}_t$ and $\boldsymbol{f}_t$. Likewise, the output gate $\boldsymbol{o}_t$ controls the emission of the memory value from the LSTM cell that learns how much of the memory cell to transfer to the hidden state $\boldsymbol{h}_t$. The depth of LSTM can be easily extended by stacking LSTMs on top of each other, where the hidden state of the LSTM in layer $l-1$ as the input to the LSTM in layer l.

LSTM can connect previous information to the present task, which can be used to extract temporal features among different frames for video scene classification. In this paper, two layers of LSTMs are stacked on the top of each other to encode temporal features for SVSC.

Finally, spatial features based on reprehensive frames and temporal encoded features by LSTM are stacked for fusion. A softmax layer is stacked on the fusion layer to predict scores for the task of SVSC.

3 Experimental setup and results

There is lack of a standard dataset for SVSC. We first build a satellite video dataset for SVSC. The proposed framework is then tested on this dataset. The description of this dataset in Subsection 3.1 and analysis of experimental results in Subsection 3.3 will be illustrated.

3.1 Dataset for satellite video scene classification

The video data is mainly collected from Jilin-1 satellite constellations, and remaining is from UrtheCast aboard International Space Station (ISS), which is public in 2016 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest [40]. These remote sensing videos provide dynamically large-scale Earth observation. The original videos last about 30∼90 s with about 1 m spatial resolution. In order to build a standard dataset for remote sensing video classification, original videos are segmented into many small video blocks with size of $64 \times 256 \times 256 \times 3$, separately denoting number of frames, width, height

Figure 7 (Color online) Samples of satellite video scenes with 8 scene classes, including airplane, runway, bridge, harbor, highway, intersection, overpass and parking lot. Four samples per each scene are shown with the first frame.

Table 1 Illustration of the built satellite video dataset

Sample statistics including overlapping areas					Sample statistics excluding overlapping areas				
Sample	Train	Test	Total	Test proportion	Sample	Train	Test	Total	Test proportion
Harbor	862	294	1156	0.25	Harbor	121	60	181	0.33
Intersection	734	184	918	0.2	Intersection	98	18	116	0.16
Highway	726	184	910	0.2	Highway	142	70	212	0.33
Bridge	388	167	555	0.3	Bridge	60	35	95	0.37
Overpass	589	162	751	0.22	Overpass	119	54	173	0.31
Airplane	788	200	988	0.2	Airplane	89	37	126	0.29
Runway	809	171	980	0.17	Runway	93	53	146	0.36
Parking lot	734	217	951	0.23	Parking lot	84	44	128	0.34
Total number	5630	1579	7209	0.22	Total number	806	371	1177	0.32

and color channels of videos. The data set is made up of the following 8 scene classes: Airplane, Runway, Bridge, Harbor, Highway, Intersection, Overpass and Parking lot. Some samples of video scenes with the first frame are shown in Figure 7. The basic principle to partition training and test dataset is to keep videos from the same region separated in training and test phases. More details of the dataset are listed in Table 1.

In addition, these video blocks are also labelled for SVSC considering moving targets. The criteria of annotation is whether videos contain moving targets. The moving targets are related with transportation, such as moving vehicles, airplanes and ships. Non-motion related patterns of significance are not taken into consideration, such as flicker (caused by noise or illumination intensity change), waves and rapid changes from overexposure sates to normal states. Finally, in the training dataset, 3840 videos have moving targets, and remaining 1790 videos have no moving targets. In the total training videos, the proportion of the satellite videos with moving target is 0.68. In the test dataset, the number of satellite videos with moving targets is 989, and the proportion of satellite videos with moving targets is 0.63.

Considering different impacting factors of video data quality and complexity, there are some challenges of this dataset for SVSC.

(1) The limit of platform and imager. Owing to the limit of platform height and the system of imagers, satellite videos have low resolution and low quality. A risk of overexposure appears during imaging

exposure, such as water bodies and smooth metal surfaces. And video cameras are adopted a pitching imaging means, which causes severe video stability problems and motion blur.

(2) Occlusion phenomena. Owing to bird's view, interesting scenes can be partly blocked by high buildings and plants in different degrees. In satellite videos, moving targets are very small compared with the size of observation scenes, where a vehicle is generally composed of several to dozens of pixels. It will lead to the difficulty for representation of moving targets, which will be easily misidentified as noise or blocked by other objects or missed with background of low contrast density.

(3) Classes complexity. The intra-class diversity is large. For the same class, different area textures, weather conditions, viewpoints and camera poses, and cluttered backgrounds will cause large variations in different videos. The inter-class diversity is small. Video blocks from same original videos always have same imaging conditions and similar color textures, which lead difficulty separating different classes.

(4) Small dataset. Labelling all videos is time-consuming and labor-intensive. In this dataset, collection of satellite remote sensing video data is finally formed into a small dataset. The small dataset with large complexity brings learning difficulty for SVSC.

3.2 Experimental setup

For SVSC, there are two-stage steps for spatial feature extraction and motion representation in our proposed framework. The spatial features are extracted based on representative frames by VGG-Net. For motion representation, the Se-TPC is first used for representing motion information of moving targets, followed by the VGG-Net and two-layer LSTM for temporal feature encoding. Finally spatial features from the first stage and motion representation from the second stage are stacked for feature fusion. A softmax layer with 8 nodes is the last (prediction) layer for SVSC. The input of the second stage is 16 Se-TPC maps, which is the transformation of two consecutive frames and uniformly sampled every four maps from satellite videos. In the two-layer LSTM, there are 1024 hidden units in the first layer, and 512 hidden units in the second layer.

In the training phase, the weights in the VGG-Net of the first stage for spatial feature extraction, two-layer LSTM and the fullly connected layer of fusion phase are needed to learn. The weights in the VGG-Net of the second stage for motion representation are directly transferred as fixed weights. The proposed network is trained by stochastic gradient descent (SGD) with cross entropy as loss function. The learning rate and momentum are set to 0.001 and 0.9. The mini-batch is 32. And the strategy of dropout is used in LSTMs with 0.5.

The proposed method is compared with several state-of-the-art methods for video scene classification including C3D [26], Two-stream CNN [27], and T-ResNet [32]. Both the fine-tuned VGG-Net and VGG-net as a feature extractor followed by support vector machine (SVM) are used for comparison. Implemented details of comparative methods are listed as follows.

(1) C3D [26] is a spatiotemporal network, which captures temporal aspects of the data while maintaining spatial information. It achieves good performance for dynamic scene classification. The pre-trained C3D network is fine-tuned on the built satellite video dataset with 0.001 learning rate. As in [26], the input of C3D is 16-frame clips, where one satellite video has 4 non-overlapping clips in our built dataset. In the test phase, the probability of 4 clips in a video is averaged as the video classification results.

(2) Two-stream CNN [27] is a two-stream CNN architecture, which incorporates spatial and temporal networks. The temporal stream uses a stack of 10 optical flow frames as input, where optical flows are extracted by a classical algorithm [41]. The spatial stream uses a random frame of a video as input. The pre-trained networks of two streams are downloaded from authors' website. The two-stream features from the last fully connected layers are stacked and formed into an 8192-dimensional descriptor. The SVM classifier is finally trained for SVSC. The regularization parameters are determined by cross validation with the range of $[10^{-5}, 10^{-4}, \ldots, 1, \ldots, 10^{4}, 10^{5}]$.

(3) T-ResNet [32] is a fully convolutional architecture in spatiotemporal orientation, which is based on temporal residual units. T-ResNet uses 16 frames from each satellite video, which randomly samples the starting frame with 2 temporal stride. During training phase, 0.2 proportion of videos in training dataset

Table 2 Results of SVSC with specific video frame (%)[a)]

Class	CNN+SVM			Fine-tuned VGG-Net		
	The 1st frame	The last frame	The representative frame	The 1st frame	The last frame	The representative frame
Harbor	95.92	**98.98**	**98.98**	94.56	87.07	75.17
Intersection	15.76	22.83	19.57	33.70	54.89	**68.48**
Highway	**35.33**	23.91	30.98	15.76	33.70	23.91
Bridge	0.00	0.00	0.00	0.60	**7.19**	1.80
Overpass	96.91	87.04	96.30	**100.00**	**100.00**	**100.00**
Airplane	80.50	78.00	82.00	90.50	74.50	**98.00**
Runway	81.29	**95.91**	94.15	87.13	66.67	78.95
Parking lot	59.45	66.36	56.68	58.53	**92.63**	88.02
AA	58.14	59.13	59.83	60.10	64.58	**66.79**
OA	60.92	62.19	62.57	62.63	66.94	**68.27**

a) The best results are in bold.

are used as validation dataset. The learning rate is 0.01 and decreases it by an order of magnitude after the validation error saturates. The mini-batch is set to 128 in the training phase. Compared with the original setup in [32], no data augmentation is used for SVSC.

(4) Fine-tuned VGG-Net. The pre-trained weights of VGG-Net are as initial weights. The last fully connected nodes are reset as 8 nodes instead of 1000 nodes according to our classification task. The network is trained on our video dataset by fine-tuning weights with a low learning rate (0.001) and a mini-batch size of 32. In addition, dropout is set as 0.5 for the first fully layers to avoid overfitting.

(5) VGG-Net + SVM. The pre-trained VGG-Net is as a feature extractor. The features from the second fully connected layer are used for SVM classification. The regularization parameters are determined by cross validation with the range of $[10^{-5}, 10^{-4}, \ldots, 1, \ldots, 10^4, 10^5]$.

3.3 SVSC experiments

3.3.1 *SVSC with specified video frame*

Firstly, in terms of specified video frame classification, the classification results of fine-tuned CNN and CNN+SVM are analyzed. The SVSC results with specified video frame are shown in Table 2 on the built satellite video scene dataset. As shown in Table 2, the classification performance using representative frames is better than the first and the last input frames according to results of three kinds of input frames. The reason is that the representative frames with low blur values and high spatial activities guarantee abundant spatial information and less motion blur. The spatial features can be better extracted with large discrimination by CNN, which are beneficial for SVSC.

The confusion matrices of all experiments in this part are shown in Figure 8. It can be seen that video scene of overpass is the easiest classified class, which is basically correctly classified by different methods. However, video scenes of bridge and highway are not easily to distinguish. If there is water in river, the scene of bridge has similar textures with harbor scene, especially for harbor without mooring many ships. If water of river dries up in some seasons, the river course under the bridges is similar with textures of road, which will lead to misclassification of bridge into overpass. For the same reason, if the river course is not obvious with cover of plants, it will lead to confusion of the scene of highway.

3.3.2 *SVSC by exploiting spatio-temporal information*

This part exploits spatio-temporal information for SVSC, where temporal information is incorporated for classification besides spatial information. Three state-of-the-art methods of video scene classification are used for comparison, including C3D [26], Two-stream CNN [27] and T-ResNet [32]. Comparing classification results with C3D, Two-stream CNN, and T-ResNet, our proposed method has higher overall accuracy (OA) and average accuracy (AA) as shown in Table 3. In our proposed method, temporal principal components are used for motion representation, which are effective to handle with information of moving

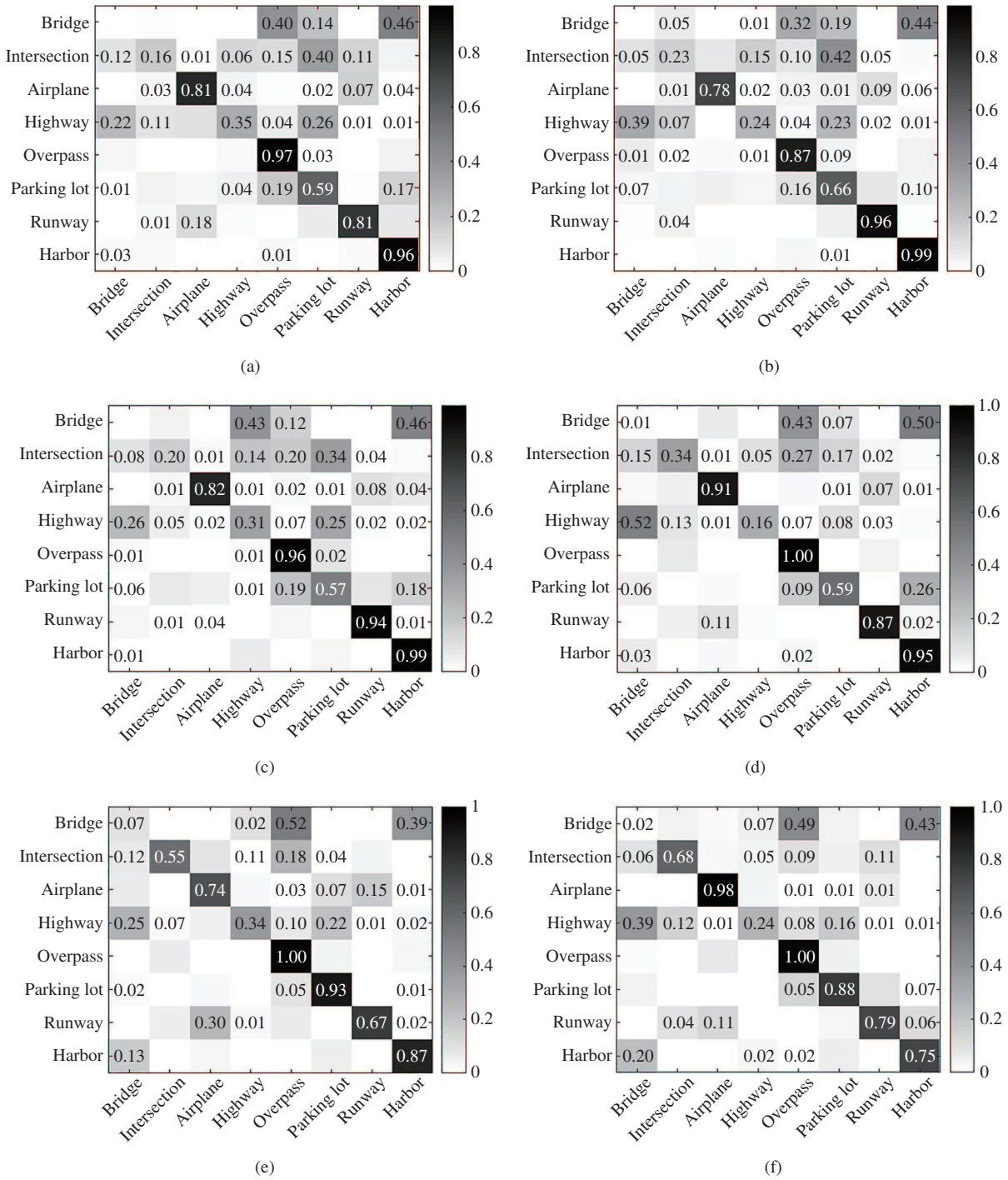


Figure 8 Normalized confusion matrices with specific video frame for SVSC. (a) The first frame with CNN+SVM; (b) the last frame with CNN+SVM; (c) the representative frame with CNN+SVM; (d) the first frame with fine-tuned VGG-Net; (e) the last frame with fine-tuned VGG-Net; and (f) the representative frame with fine-tuned VGG-Net. The rows and columns of the matrix respectively denote the predicted and actual labels.

targets in satellite videos. The temporal coding by VGG-Net and LSTM can further extract temporal features for SVSC. However, optical flows used in Two-stream CNN, are interfered by non-motion related patterns of significance, such as flicker (caused by noise or illumination intensity change), waves and rapid changes from overexposure sates to normal states. It will reduce classification performance. Both 3D convolutions in C3D and temporal units in T-ResNet belong to implicitly exploit temporal features. In terms of specific characteristics in satellite videos, classification results from both C3D and T-ResNet are

Table 3 Results of SVSC by exploiting spatio-temporal information (%)[a)]

	C3D [26]	Two-stream CNN [27]	T-ResNet [32]	Proposed method
Harbor	90.48	85.03	**99.32**	98.30
Intersection	57.07	42.39	40.22	**80.98**
Highway	5.44	30.43	5.43	**32.07**
Bridge	1.20	0.00	**4.79**	0.00
Overpass	**100.00**	91.36	**100.00**	0.94
Airplane	**100.00**	60.00	79.50	94.44
Runway	**89.47**	64.33	73.10	69.59
Parking lot	7.37	10.60	59.45	**96.77**
AA	55.77	48.02	57.83	**70.83**
OA	57.25	49.72	60.67	**73.97**

a) The best results are in bold.

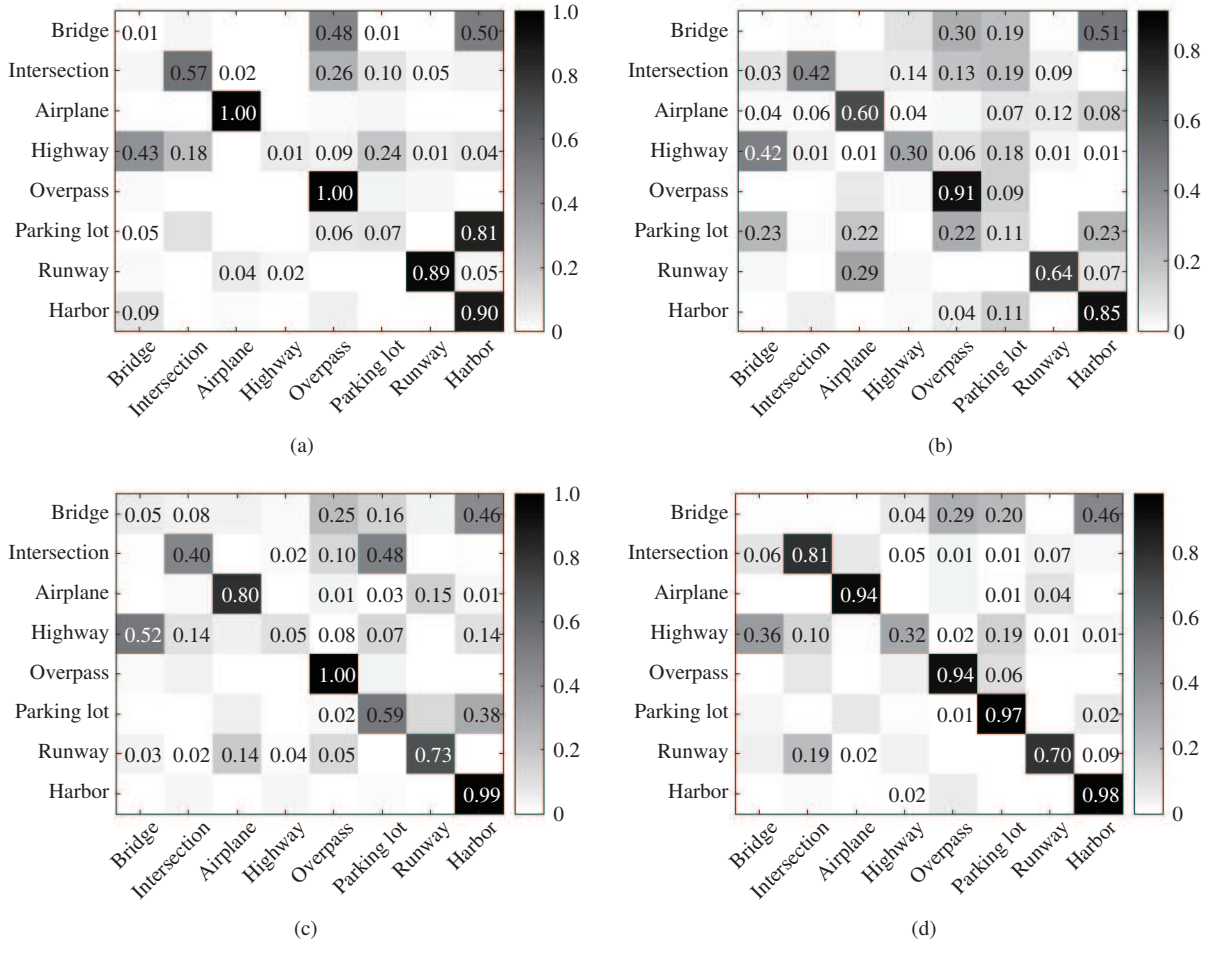


Figure 9 Normalized confusion matrices by exploiting spatio-temporal information for SVSC. (a) C3D [26]; (b) two-stream CNN [27]; (c) T-ResNet [32]; (d) our proposed method. The rows and columns of the matrix respectively denote the predicted and actual labels.

not ideal. There are small moving targets and most part of unchanged scenes in our dataset, which causes the smaller receptive field for moving targets. Therefore, both C3D and T-ResNet have limited capacities to extract temporal features for SVSC. For instance, both C3D and T-ResNet have lower classification accuracy for highway scene, where temporal information is badly extracted as shown in Table 3.

The confusion matrices of all experiments in this part are shown in Figure 9. It can be seen that video scene of overpass is still the easiest classified class, which is basically correctly classified by different

Table 4 Results of SVSC considering moving targets (%)

	Our proposed method		Movement track length	
	Frame difference	Se-TPC	Frame difference	GMM
OA	84.61	86.76	83.85	77.33

methods. However, video scenes of bridge and highway are not easily to distinguish. The bridge scenes are the most misclassified as the harbor scene, owing to similar textures with harbor scene. Compared with other methods, our proposed method performs better, especially for intersection scene and parking lot scene.

3.3.3 *SVSC considering moving targets*

To verify the effectiveness of motion representation in our proposed method, several compared experiments are implemented for SVSC considering moving targets. The criteria of classification considering moving targets are whether videos contain moving targets. The moving targets are related with transportation, such as moving vehicles, airplanes and ships. Non-motion related patterns of significance are not taken into consideration, such as flicker (caused by noise or illumination intensity change), waves, and rapid changes from overexposure sates to normal states.

The compared methods include frame difference and an improved adaptive Gaussian mixture modeling (GMM) [42]. First, in our proposed framework for temporal feature extraction, frame difference results replace the input Se-TPCs to represent motion information of moving targets in satellite videos. The results are shown in Table 4. Se-TPC has better classification performance than frame difference. Owing to low imaging quality, satellite videos are full of dense noise. Frame difference is computed by difference between consecutive frames, which is unavoidably impacted by different noise intensities and camera motion in consecutive frames. Instead, PCA has robustness to extract main information even with noise. The background scene is basically unchanged, and the size of moving targets is bigger than noise. Therefore, motion information of moving targets can be better obtained with a certain degree of noise than the method based on frame difference.

Second, the accumulative track lengths of multiple moving targets are also taken into consideration, where if a track length is bigger than threshold, this satellite video contains moving targets. Moving targets of each frame are preliminarily detected by frame difference and an improved GMM. GMM is computed using OpenCV toolbox. Then trajectories are extracted by Hungarian algorithm. In terms of empirical analysis in training data set, the threshold is set as 5 pixels. According to Table 4, modeling motion information by LSTM has obtained better classification performance than simple accumulative moving track length using simple detection methods. Owing to the occlusion problem caused by buildings and plants and low contrast intensity with background, moving targets are easily missed for motion representation. In addition, owing to strong reflection of water bodies and metal surfaces (e.g., some metal roofs of building and vehicles), the initial overexposure phenomenon is obvious, which is easily caused misjudgment to moving targets.

4 Conclusion

In this paper, a dataset of satellite videos is built for satellite video scene classification. The spatial resolution of satellites video is about 1 m, which can identity fine scenes on Earth. The occlusion problems and scene complexity bring big challenges for SVSC. In our proposed method, fine-tuned VGG-Net is transferred to extract spatial features from representative frames. Representative frames guarantee abundant spatial information and less motion blur, which have better classification performance than classification based on some specific frames. Temporal component principles are extracted as the motion information of moving targets, which effectively represent small moving targets. Then temporal features are further encoded by VGG-Net and two-layer LSTM. The temporal features have better ability for SVSC considering moving targets, which are determined by whether videos contain moving targets. For

SVSC, compared with the state-of-the-art video classification methods, our proposed method based on spatial features and temporal features, has higher classification accuracy. However, occlusion problem, low imaging quality, and overexposure bring misclassification for some specific classes, e.g., bridge scene. Experimental results show the effectiveness of our proposed framework.

However, these current experimental results are preliminarily applied for satellite video scene classification, so more experiments and analysis will be done in the future. In addition, there is a lot of work for further analyzing satellite videos as the development of remote sensing techniques, e.g., detection and tracking of moving targets, action pattern understanding in large-scale satellite videos.

Acknowledgements This work was supported by National Natural Science Foundation of Key International Cooperation (Grant No. 61720106002), Key Research and Development Project of Ministry of Science and Technology (Grant No. 2017YFC1405100), National Natural Science Foundation of China (Grant No. 61901141), and Fundamental Research Funds for the Central Universities (Grant No. HIT.HSRIF.2020010). The authors would like to thank the IEEE GRSS Image Analysis and Data Fusion Technical Committee for providing Urthecast satellite videos.

References

1 Yan C, Xie H, Chen J, et al. A fast Uyghur text detector for complex background images. IEEE Trans Multimedia, 2018, 20: 3389–3398
2 Wang Q Q, Huang Y, Jia W J, et al. FACLSTM: ConvLSTM with focused attention for scene text recognition. Sci China Inf Sci, 2020, 63: 120103
3 Zhao J P, Guo W W, Zhang Z H, et al. A coupled convolutional neural network for small and densely clustered ship detection in SAR images. Sci China Inf Sci, 2019, 62: 042301
4 Marszalek M, Laptev I, Schmid C. Actions in context. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, 2009. 2929–2936
5 Yan C, Tu Y, Wang X, et al. STAT: spatial-temporal attention mechanism for video captioning. IEEE Trans Multimedia, 2020, 22: 229–241
6 Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), New York, 2006. 2169–2178
7 Sánchez J, Perronnin F, Mensink T, et al. Image classification with the fisher vector: theory and practice. Int J Comput Vis, 2013, 105: 222–245
8 Cheriyadat A M. Unsupervised feature learning for aerial scene classification. IEEE Trans Geosci Remote Sens, 2014, 52: 439–451
9 Yan C, Li L, Zhang C, et al. Cross-modality bridging and knowledge transferring for image understanding. IEEE Trans Multimedia, 2019, 21: 2675–2685
10 Othman E, Bazi Y, Alajlan N, et al. Using convolutional features and a sparse autoencoder for land-use scene classification. Int J Remote Sens, 2016, 37: 2149–2167
11 Otavio A B P, Nogueira K, dos Santos J A. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, 2015. 44–51
12 Hu F, Xia G S, Hu J, et al. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens, 2015, 7: 14680–14707
13 Chaib S, Liu H, Gu Y, et al. Deep feature fusion for VHR remote sensing scene classification. IEEE Trans Geosci Remote Sens, 2017, 55: 4775–4784
14 Li E, Xia J, Du P, et al. Integrating multilayer features of convolutional neural networks for remote sensing scene classification. IEEE Trans Geosci Remote Sens, 2017, 55: 5653–5665
15 He N, Fang L, Li S, et al. Remote sensing scene classification using multilayer stacked covariance pooling. IEEE Trans Geosci Remote Sens, 2018, 56: 6899–6910
16 Yi S, Pavlovic V. Spatio-temporal context modeling for BoW-based video classification. In: Proceedings of IEEE International Conference on Computer Vision Workshops (ICCVW), Sydney, 2013. 779–786
17 Zhao G Y, Ahonen T, Matas J, et al. Rotation-invariant image and video description with local binary pattern features. IEEE Trans Image Process, 2012, 21: 1465–1477
18 Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM International Conference on Multimedia (ACMMM), Augsburg, 2007. 357–360
19 Derpanis K G, Lecce M, Daniilidis K, et al. Dynamic scene understanding: the role of orientation features in space and time in scene classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, 2012. 1306–1313
20 Wang H, Ullah M M, Klaser A, et al. Evaluation of local spatio-temporal features for action recognition. In: Proceedings of British Machine Vision Conference (BMVC), London, 2009. 1–11
21 Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition. Int J Comput Vis, 2013, 103: 60–79

22 Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), Sydney, 2013. 3551–3558

23 Karpathy A, Toderici G, Shetty S, et al. Large scale video classification with convolutional neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, 2014. 1725–1732

24 Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, 2018. 6546–6555

25 Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. In: Proceedings of IEEE International Conference on Computer Vision Workshop (ICCVW), Venice, 2017. 3154–3160

26 Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), Santiago, 2015. 4489–4497

27 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), Quebec, 2014. 568–576

28 Donahue J, Hendricks L A, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 677–691

29 Srivastava N, Mansimov E, Salakhutdinov R. Unsupervised learning of video representations using LSTMs. In: Proceedings of International Conference on Machine Learning (ICML), Lille, 2015. 843–852

30 Ng J Y, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: deep networks for video classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, 2015. 4694–4702

31 Zhu L, Xu Z, Yang Y. Bidirectional multirate reconstruction for temporal modeling in videos. In: Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017. 1339–1348

32 Feichtenhofer C, Pinz A, Wildes R P. Temporal residual networks for dynamic scene recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017. 7435–7444

33 Simonyan K, Zisserman A. Very deep convolutional networks for large scale image recognition. In: Proceedings of International Conference on Learning Representations (ICLR), San Diego, 2015. 1–14

34 Liu T M, Zhang H J, Qi F H. A novel video key-frame-extraction algorithm based on perceived motion energy model. IEEE Trans Circ Syst Video Technol, 2003, 13: 1006–1013

35 Sze K W, Lam K M, Qiu G P. A new key frame representation for video segment retrieval. IEEE Trans Circ Syst Video Technol, 2005, 15: 1148–1155

36 Dufaux F. Key frame selection to represent a video. In: Proceedings of International Conference on Image Processing (ICIP), Vancouver, 2000. 275–278

37 Crete F, Dolmiere T, Ladret P, et al. The blur effect: perception and estimation with a new no-reference perceptual blur metric. In: Proceedings of SPIE, 2007. 64920I

38 Sahouria E, Zakhor A. Content analysis of video using principal components. IEEE Trans Circ Syst Video Technol, 1999, 9: 1290–1298

39 Xia G S, Hu J, Hu F, et al. AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans Geosci Remote Sens, 2017, 55: 3965–3981

40 Tuia D, Moser G, Le Saux B. 2016 IEEE GRSS data fusion contest: very high temporal resolution from space technical committees. IEEE Geosci Remote Sens Mag, 2016, 4: 46–48

41 Farneback G. Two-frame motion estimation based on polynomial expansion. In: Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA), 2003. 363–370

42 KaewTraKulPong P, Bowden R. An improved adaptive background mixture model for real-time tracking with shadow detection. In: Proceedings of the 2nd European Workshop on Advanced Video Based Surveillance System, Boston, 2002. 135–144