J. Vis. Commun. Image R. 80 (2021) 103267

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Multispectral background subtraction with deep learning ☆

Rongrong Liu *, Yassine Ruichek, Mohammed El Bagdouri

Connaissance et Intelligence Artificielle Distribuées (CIAD), University Bourgogne Franche-Comté, UTBM, 12 Rue Thierry Mieg, 90000, Belfort, France

ARTICLE INFO

MSC:
68T45
68U10
94A08
94A13

Keywords:
Background subtraction
Multispectral images
Deep learning
Convolutional neural networks

ABSTRACT

In this paper, we follow the trend of deep learning and make an attempt to investigate the potential benefit of using multispectral images via convolutional neural networks for background subtraction task. The major contributions of this work lie in two aspects, based on the impressive algorithm FgSegNet_v2. Firstly, we extract three channels out of the seven of the FluxData FD-1665 multispectral dataset to match the number of input channels of the VGG16 deep model. Some combinations of three-channel based multispectral images perform better than RGB images. Secondly, a new convolutional encoder is designed to use all the multispectral channels available to further explore the information of multispectral images. The results outperform the RGB images and also other approaches using the same multispectral dataset.

1. Introduction

Background subtraction is also called change detection [1], foreground detection [2] and foreground–background segmentation [3]. As the name suggests, it aims to detect foreground regions that are in motion from background of a video sequence [4] and it plays an important and perquisite role in computer vision due to its potential applications [5], such as intelligent surveillance, traffic monitoring and industrial machine vision.

The first task of background subtraction is to establish a solid background model [6]. Then, this model is used as a reference to compare with the current image to label pixels as background or foreground pixels. Background subtraction is a well studied field. Therefore there exists a vast number of algorithms for this purpose [1]. Detailed overviews of background subtraction methods are available in [7, 8]. Representative conventional methods include Mixture of Gaussian (MoG) [9], also called Gaussian Mixture Models (GMM), Kernel Density Estimation (KDE) [10], Codebook [11] and ViBe [12].

In this decade, deep learning based on the work of Yann LeCun et al. in 1989 [13], has revolutionized computer vision, and deep features obtained from Convolutional Neural Networks (ConvNets also called CNNs) have been shown as powerful and effective image representations for various computer vision tasks such as object classification [14–17], object detection [18–22] semantic segmentation [23–25]. Recently, inspired by the impressive achievement of deep learning, background subtraction based on ConvNets shows great success and is now becoming a hot research topic due to its high precision [26,27].

The first attempt to apply ConvNets for background subtraction problem was conducted by Braham and Van Droogenbroeck in 2016 [28]. Since then, numerous supervised-based deep learning papers [1, 26,27,29–33] have been published in the field of background subtraction. Currently, the top background subtraction methods in the large-scale dataset CDnet2012 [34] and its extension CDnet 2014 [35] are based on ConvNets with a large gap of performance in comparison to conventional approaches. Among all these supervised-based deep learning methods, the method called FgSegNet _v2 [27] outperforms state-of-the-art approaches.

Moreover, the majority of the current background methods in this research community are focused on visible images or Red–Green–Blue (RGB). With the rise of different sensors, multi-modal foreground detection, which integrates multiple complementary data like visible and thermal infrared sources, has received more and more attention recently [36]. Compared with visible images, background subtraction using multispectral images can be more interesting because of the better spectral resolution. Thanks to the recent advances in technology, new products such as the FD-1665 Multispectral Cameras from FluxData are commercially available to offer the possibility to record multispectral images of more than three spectral channels in the visible and near infrared(NIR) part of the spectrum simultaneously [37].

To our knowledge, the FluxData FD-1665 dataset is the only public real multispectral image background subtraction dataset. Most public image datasets built for background subtraction, or change detection, such as the well-known Wallflower dataset [38], the Stuttgart Artificial

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.

E-mail addresses: rongrong.liu@utbm.fr (R. Liu), yassine.ruichek@utbm.fr (Y. Ruichek), mohammed.el-bagdouri@utbm.fr (M. El Bagdouri).

https://doi.org/10.1016/j.jvcir.2021.103267
Received 8 September 2020; Received in revised form 29 June 2021; Accepted 7 August 2021
Available online 14 August 2021

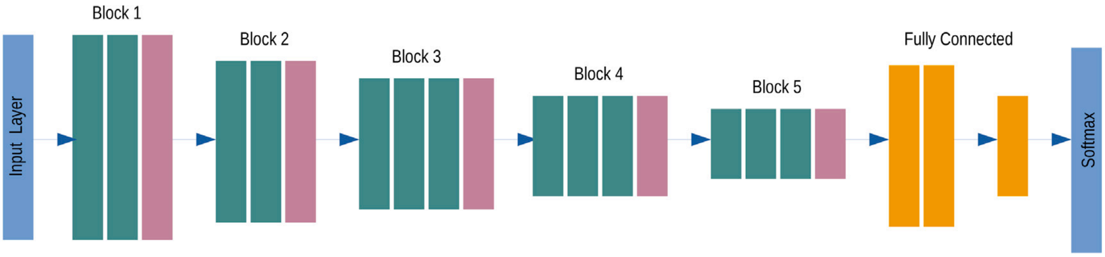

Fig. 1. Architecture of VGG16.

Background Subtraction (SABS) dataset [39] and CDnet [34,35], are based on visible spectral images or still recombined images. For example, the Grayscale-Thermal Foreground Detection (GTFD) dataset [40] provides pairs of grayscale and thermal frames to investigate the fusion methods of thermal and grayscale data for effective foreground detection.

According to the above observations and motivated by the impressive accuracy of FgSegNet_v2 for foreground segmentation with RGB images, we tried to explore the benefits of multispectral images in background subtraction with deep learning based on this deep model. In order to achieve this goal, we first extracted all the possible combinations of three channels from multispectral images to match the input of the FgSegNet_v2 deep model, aiming to investigate the possible improvements against RGB. Then a new convolutional encoder was proposed to utilize all the multispectral channels available to further study their potential benefits via deep features learned with ConvNets for the background subtraction task.

The rest parts of this paper are organized as follows. Section 2 briefly discusses the background and related works. The proposed attempts utilizing the ConvNets-based multispectral background subtraction methods are explained in Section 3. Section 4 describes the experimental evaluation including the used dataset, the training details and the obtained results compared with other approaches using the same dataset. In Section 5, we conclude our paper and provide some future works.

2. Related works

2.1. Deep learning

Deep models can be referred to as neural networks with deep structures. The concept of neural networks is not something new and can date back to 1940s [41]. The original intention was to simulate the human brain system to solve general learning problems in a principled way. Since then this decades-old scientific discovery started its long journey to innovate the entire academic community.

There are some remarkable technical breakthroughs and significant advances in the design of network structures and training strategies, including but not limited to: [42] proposed back-propagation algorithm in 1980s; [43] came up with a new type of non-linearity, namely the widely used Rectified Linear Unit (ReLU); [44] proposed a new weight initialization; [45] proposed the Max-pooling instead of average sub-sampling; with dropout [46] and data augmentation, the overfitting problem in training could be relieved; with batch normalization (BN) [47], the training of very deep neural networks becomes quite efficient.

It is all these continuous efforts, together with the emergence of large scale annotated training data, such as ImageNet [48] and the fast development of high performance parallel computing systems, such as graphics processing units (GPUs) that prosper deep learning nowadays [49].

2.2. Convolutional neural networks

ConvNets are the most representative models of deep learning and are designed to process data that come in the form of multiple arrays, such as a color image composed of three 2D arrays containing pixel intensities in the three color channels [50]. ConvNets-based network architectures now dominate the field of computer vision.

Although ConvNets were invented in the 1980s, the breakthrough was made on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)-2012 by Krizhevsky et al. [14], who applied the deep convolutional networks to a dataset of 1.2 million high-resolution images and achieved excellent performance, almost halving the error rates of the best competing approaches. This network is called AlexNet, named after Alex Krizhevsky, the first author of this breakthrough. The results of AlexNet, for the first time, show that a large convolutional neural network is capable of achieving recordbreaking results on a highly challenging dataset using purely supervised learning.

Since AlexNet, even larger and deeper networks have been proposed. These models include the VGG networks [15], which make use of a number of repeating blocks of elements; the network in network (NiN) [51], which convolves whole neural networks patch-wise over inputs; the GoogLeNet [16] and its higher version [47,52] and [53], which make use of networks with parallel concatenations; residual networks (ResNet) [17] which are currently the most popular go-to architecture today, and densely connected networks (DenseNet) [54], which are expensive to compute but have set some recent benchmarks [55]. These architectures are now the base models upon which an enormous amount of research and projects are built and have been applied with great success in the computer vision community.

2.3. VGG16 network

The VGG networks are CNN designed by Simonyan and Zisserman [15]. They were originally proposed for the ImageNet Large Scale Visual Recognition Competition (ILSVRC-2014) by the Visual Geometry Group, where the name of this set of models comes from.

There are 6 kinds of ConvNet configurations included in the VGG networks. All configurations follow the generic design in architecture and differ only in the depth. As a typical and popular ConvNet architecture, VGG16 has been fine-tuned for many other tasks, such as object detection [21], semantic segmentation [29,56] and so on.

Fig. 1 shows the VGG16 architecture. Like earlier AlexNet [14], VGG16 can be partitioned into two parts: the first one consisting mostly of convolutional and pooling layers, and the second one consisting of fully connected layers. The idea of block is used in the first part. The VGG16 has 5 convolutional blocks, among which the first two have two convolutional layers each and the latter three contain three convolutional layers each.

Table 1 illustrates the details of VGG16 configuration, including the type, the kernel size, the number of channels and the output shape for each hidden layer. In detail, one VGG block consists of a sequence of convolutional layers, performing 3 × 3 convolutions with stride 1 and pad 1, followed by a maxpooling layer for spatial downsampling, which

Table 1
VGG16 network configuration.

Block	Layer type	Kernel size	Number of channels	Output shape
	Input	–	–	224 × 224 × 3
1	Convolution	3 × 3	64	224 × 224 × 64
	Convolution	3 × 3	64	224 × 224 × 64
	Maxpooling	2 × 2		112 × 112 × 64
2	Convolution	3 × 3	128	112 × 112 × 128
	Convolution	3 × 3	64	112 × 112 × 128
	Maxpooling	2 × 2		56 × 56 × 128
3	Convolution	3 × 3	256	56 × 56 × 256
	Convolution	3 × 3	256	56 × 56 × 256
	Convolution	3 × 3	256	56 × 56 × 256
	Maxpooling	2 × 2		28 × 28 × 256
4	Convolution	3 × 3	512	28 × 28 × 512
	Convolution	3 × 3	512	28 × 28 × 512
	Convolution	3 × 3	512	28 × 28 × 512
	Maxpooling	2 × 2		14 × 14 × 512
5	Convolution	3 × 3	512	14 × 14 × 512
	Convolution	3 × 3	512	14 × 14 × 512
	Convolution	3 × 3	512	14 × 14 × 512
	Maxpooling	2 × 2		7 × 7 × 512
	FullyConnected		4096	1 × 1 × 4096
	FullyConnected		4096	1 × 1 × 4096
	FullyConnected		1000	1 × 1 × 1000

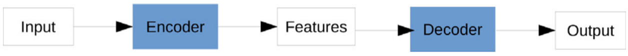

Fig. 2. Encoder-decoder architecture.

performs 2 × 2 maxpooling with stride 2. Thus the resolution is halved after each block.

We can also see in Table 1, the first block has 64 output channels and each subsequent block doubles the number of output channels, until that number reaches 512 [55]. The input size of pictures for VGG16 is 224 × 224 × 3 and 3 stands for the three spectral channels of RGB. The output size for each layer is also listed to better understand the structure of this model.

2.4. Encoder–decoder architecture

The encoder–decoder architecture is a practical neural network design pattern and it is applied in many image classification network, such as AlexNet [14], VGG networks [15], GoogLeNet [16] and Resnet [17]. In this architecture, the network is partitioned into two parts, the encoder and the decoder.

As Fig. 2 shows, the input, such as an image patch, is fed to the encoder which produces features. The decoder module then converts the features into prediction results for a specific purpose. Meantime, the error is measured. The encoder and decoder are parameterized functions that are trained to minimize the average error [57].

2.5. Background subtraction methods based on multispectral images

In this subsection, we will introduce several background subtraction approaches with multispectral images. They have been tested on the same dataset using all the seven channels and will be later used for quantitative comparison with the convolutional approach proposed in this paper.

2.5.1. Mahalanobis distance

The first method we will review here is a straightforward extension of a color-based background subtraction algorithm, which is introduced in the original multispectral dataset paper [37]. As it is known, background subtraction can be performed by the Formula (1), with the assumption that the observed video sequence $\boldsymbol{I}$ is made of a static background $\boldsymbol{B}$ in front of which moving objects are observed.

$$\chi_t(s) = \begin{cases} 1, & d(\boldsymbol{I}_{s,t}, \boldsymbol{B}_{s,t}) > \tau \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where τ is a threshold, χ_t is the motion label field at time t, and d is the distance between the pixel value $\boldsymbol{I}_{s,t}$ and the background model $\boldsymbol{B}_{s,t}$ at time t and location s. For a multispectral video sequence, $\boldsymbol{I}_{s,t}$ is a vector defined by $\boldsymbol{I} = [\boldsymbol{I}_1, \boldsymbol{I}_2, \dots, \boldsymbol{I}_n]$, where n stands for the number of spectral channels of multispectral frames.

If the background $\boldsymbol{B}$ can be determined by a single image free of moving objects, pixels corresponding to foreground moving objects can be detected by thresholding a distance function, such as the Euclidian distance:

$$d = \sqrt{\left(\sum_{i=1}^{k}(I^i_{s,t} - B^i_{s,t})^2\right)}. \tag{2}$$

However, it is not the case for real-life scenarios. Modeling the background $\boldsymbol{B}$ with a single image requires a rigorously fixed background void of noise and artefacts. A promising solution is to model each background pixel by a probability density function (PDF) learned over a series of training frames. In this case, the background subtraction problem becomes a PDF-thresholding issue for which a pixel with low probability is likely to correspond to a foreground moving object.

For instance, in order to account for noise, it is possible to model every background pixel with a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ stand for the average background multispectral vector and covariance matrix at pixel s and time t, respectively. In this context, the distance metric can be the Mahalanobis distance:

$$d_M = \sqrt{|\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t}|\boldsymbol{\Sigma}^{-1}_{s,t}|\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t}|^T}. \tag{3}$$

Since the covariance matrix contains large values in noisy areas and low values in more stable areas, $\boldsymbol{\Sigma}$ makes the threshold locally dependent on the amount of noise. In other words, the noisier a pixel is, the larger the temporal gradient $|\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t}|$ has to be to get the pixel labeled in motion. This makes the method significantly more flexible and robust.

Besides, In order to adjust the fact that the illumination often changes in time, the mean and covariance of each pixel can also be iteratively updated with:

$$\boldsymbol{\mu}_{s,t+1} = (1-\alpha)\boldsymbol{\mu}_{s,t} + \alpha\boldsymbol{I}_{s,t} \tag{4}$$

$$\boldsymbol{\Sigma}_{s,t+1} = (1-\alpha)\boldsymbol{\Sigma}_{s,t} + \alpha(\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t})(\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t})^T \tag{5}$$

2.5.2. Spectral angle

Instead of using a straightforward extension of the color-based Mahalanobis distance, the multispectral dataset paper [37] has also proposed another two dedicated ways to measure the similarity or dissimilarity of spectral vectors.

The first one is called spectral angle d_θ, which extracts geometric features by calculating the angle between two spectra [58] and is defined with the following formula:

$$d_\theta(\boldsymbol{I}_{s,t}, \boldsymbol{\mu}_{s,t}) = \cos^{-1}\left(\frac{\langle \boldsymbol{I}_{s,t}, \boldsymbol{\mu}_{s,t}\rangle}{|\boldsymbol{I}_{s,t}||\boldsymbol{\mu}_{s,t}|}\right), \tag{6}$$

where $\boldsymbol{I}_{s,t}$ and $\boldsymbol{\mu}_{s,t}$ are the multispectral vectors of the current image and the background model, respectively. As we can see, here spectra are considered as vectors in a k-dimensional space, which indicates this spectral distance measure is suitable for arbitrary number of multispectral channels.

With spectral angle metric, small angles mean similar vectors. Thus, another key advantage of this measure is that it is intensity invariant because the angle between two vectors is independent of the vector length. This property is very interesting for background subtraction problems when there are shadows and illumination variations.

2.5.3. Spectral information divergence

Another spectral distance measurement utilized in the dataset paper [37] is referred to as Spectral Information Divergence (SID) [59], which is also applied to determine the spectral closeness or distance between two multispectral vectors. This measure is relatively recent and is expected to be more effective than spectral angle d_θ in preserving spectral properties.

The spectral information divergence models the spectral channel-to-channel variability as a result of uncertainty caused by randomness, which is based on the Kullback–Leibler divergence to measure the discrepancy of probabilistic behaviors [59]. That is to say, it considers each pixel as a random variable and then defines the desired probability distribution by normalizing its spectral histogram to unity, which is expressed by

$$\begin{cases} P_x(i) = \dfrac{x_t(i)}{\sum_{i=1}^{n} x_t(i)} \\ P_v(i) = \dfrac{v_m(i)}{\sum_{i=1}^{n} v_m(i)} \end{cases} \tag{7}$$

where n is the number of channels. Then the spectral information divergence d_{SID} between the current spectral vector $\mathbf{x}_t$ and the background model $\mathbf{v}_m$ can be defined with

$$d_{SID}(\mathbf{x}_t, \mathbf{v}_m) = \sum_{i=1}^{n} P_x(i) log \frac{P_x(i)}{P_v(i)} + \sum_{i=1}^{n} P_v(i) log \frac{P_v(i)}{P_x(i)} \tag{8}$$

2.5.4. Online stochastic tensor decomposition

Besides, Sobral et al. [60] have proposed another mechanism based on stochastic decomposition of low-rank and sparse components for background subtraction with multispectral video sequences. In this algorithm, each multispectral image is represented as a three-dimension data cube or tensor, which can be considered as a multidimensional or N-way array.

As reviewed in [61–63], low-rank and sparse decomposition methods are based on the assumption that the uncorrupted information lies in a low dimensional subspace, where noise is sparse. This assumption holds a particular association to the task of foreground–background segmentation. To be more specific, as it is almost static and highly correlated between frames, the background is assumed to be a low dimensional subspace, where the sparse outliers usually represent the foreground objects.

Based on the pioneering tensor-based decomposition methods [64, 65], the framework of Online Stochastic Tensor Decomposition (OSTD) has been proposed in [60]. They have extended the online stochastic principal analysis optimization for multispectral images using tensor analysis, where the stochastic optimization is applied on each mode of the tensor and the individual basis is updated iteratively followed by the processing of the current frame.

2.5.5. Online one-class ensemble for feature selection

In background subtraction, the features characterize a region and can be compared against a known background model to classify it as either foreground or background. As we know, color features, edge features, stereo features, and motion features are commonly used in this field. However, the optimal features for background subtraction may be case by case.

Given the ensemble learning mechanism proposed in [66], an algorithm called Online Weighted One-Class Random Subspace (OWOC-RS) has been designed by Silva et al. in [67], which is capable of selecting suitable pixel-based features to separate the foreground regions from the background. Besides, the relative importance of each feature over time is updated with adaptive mechanism.

In order to not only increase the efficiency in terms of time and memory consumption but also the segmentation performance, an improved version has been also proposed by the same authors in [68]. The novel method is named as Superpixel-based Online Wagging One-Class Ensemble (Superpixel-OWAOC) as it adopts the superpixel idea and is based on wagging for selecting suitable individual features.

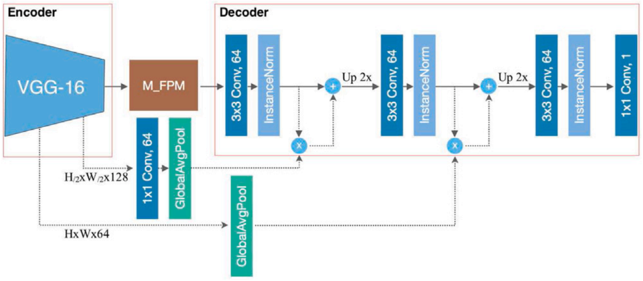

Fig. 3. FgSegNet_v2 network architecture.

3. Proposed multispectral background subtraction with deep learning

In this section, we will present the deep learning approaches we proposed for background subtraction with multispectral images. The proposed algorithms are based on FgSegNet_v2 [27], which will be first reviewed. This RGB-based model serves as an inspiration for multispectral case. Then, the proposed architectures are introduced.

3.1. FgSegNet_v2

The ranking first algorithm on the large-scale change detection dataset CDnet [34,35] is called FgSegNet_v2 [27], followed by its former version FgSegNet [26], which includes two approaches, namely FgSegNet_M and FgSegNet_S.

FgSegNet_M has proposed a triplet of encoders to extract multi-scale features and then used transposed convolution in the decoder to learn a mapping from feature space to image space. FgSegNet_S has adopted a single encoder but achieved high performance by applying several parallel dilated convolutions with different dilation rates to extract multi-scale features. As a higher version, FgSegNet_v2 has been adapted from FgSegNet_S, and reached state-of-the-art results by integrating attention mechanism to learn multi-scale features without incorporating temporal data.

FgSegNet_v2 utilizes the concept of the aforementioned encoder–decoder structure and takes advantages of the pre-trained VGG16 and the transposed convolutional neural network (TCNN). The configuration for the encoder part of the FgSegNet_v2 is shown in Table 2. It utilizes the first four blocks of VGG16 with some modifications, i.e. removing the maxpooling layer of the third and forth blocks and inserting a dropout layer after every convolutional layer in the forth block. During the training process, the first three blocks are frozen and only the modified forth block is reinitialized and finetuned.

In order to use both higher level and lower level contextual information, Lim et al. [27] have designed a Feature Pooling Module (FPM), as shown in Fig. 3. Then, these concatenated features at multiple scales are fed to the TCNN for decoding the feature maps. Finally, a threshold is applied to the resulting grayscale image to obtain a binary foreground background segmentation mask. Hereafter, we refer to the FPM also as a part of the decoder network.

3.2. Multispectral three-channel based FgSegNet_v2

The original VGG16 deep model is pretrained using conventional RGB images with three channels. Accordingly, the third dimension of the filter for the first convolutional layer is also three. In order to investigate the possible improvement of multispectral images against RGB based on FgSegNet_v2, we first extract three channels out of seven in the multispectral FluxData FD-1665 dataset [37], which will be introduced in detail later in the next section. Then, the trials with these extracted three-channel images are conducted using FgSegNet_v2 for each scene.

Table 2
FgSegNet_v2 encoder configuration.

Block	Layer type	Kernel size	Number of channels	Output shape
	Input	–	–	W ×H ×3
1	Convolution	3 × 3	64	W × H × 64
	Convolution	3 × 3	64	W × H × 64
	Maxpooling	2 × 2		W/2 × H/2 × 64
2	Convolution	3 × 3	128	W/2 × H/2 × 128
	Convolution	3 × 3	64	W/2 × 112 × H/2
	Maxpooling	2 × 2		W/4 × H/4 × 128
3	Convolution	3 × 3	256	W/4 × H/4 × 256
	Convolution	3 × 3	256	W/4 × H/4 × 256
	Convolution	3 × 3	256	W/4 × H/4 × 256
4	Convolution	3 × 3	512	W/4 × H/4 × 512
	Dropout			W/4 × H/4 × 512
	Convolution	3 × 3	512	W/4 × H/4 × 512
	Dropout			W/4 × H/4 × 512
	Convolution	3 × 3	512	W/4 × H/4 × 512
	Dropout			W/4 × H/4 × 512

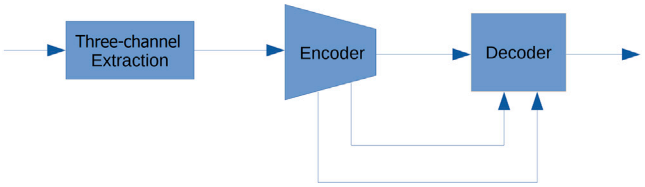

Fig. 4. Multispectral three-channel based FgSegNet_v2.

Fig. 4 illustrates the working flow of this proposed mechanism. The multispectral images are first processed through the module of three-channel extraction to produce three-channel based images, which are then fed to the encoder adapted from VGG16 and decoder network, following the same manner of FgSegNet_v2 explained above.

The number of combinations composed of three channels among seven is $C_7^3 = 35$. Thus, for each scene, we have built thirty-five independent background models, each one corresponds to a combination of three channels. After each deep model is trained with the training subset of its corresponding combination of channels, the testing subset of the same combination of channels is used to evaluate the performance of this background model.

3.3. Proposed convolutional encoder for all multispectral channels

As we mentioned in the last section, the third dimension value of the first filter of the original VGG16 is three, as it has been pretrained with three-channel based RGB images. If we simultaneously feed the multispectral images with more than three channels to the deep model of FgSegNet_v2, which adopts the first four blocks of VGG16 as the encoder, only the first three channels are really processed, while the others are ignored. Therefore, we cannot utilize the FgSegNet_v2 model directly for multispectral images with more than three channels.

In order to further explore the benefits of multispectral images, we have proposed a new convolutional encoder to extract the relevant deep features from the given multispectral-groundtruth pair with images consisting any arbitrary number of channels. Table 3 illustrates the configuration of the proposed encoder for multispectral images, where the number of input is seven, as the FluxData FD-1665 dataset we use contains seven channels. Including the trainable Block 4 from the FgSegNet_v2, the proposed encoder consists of two more convolutional layers, both of which are followed by a maxpooling layer in order to match the size of the inputs for the decoder.

Fig. 5 shows the architecture of background subtraction for multispectral images with the proposed convolutional encoder. Following the idea of FgSegNet_v2 that feeds different levels of features to the decoder, the low level feature coefficients vectors after each convolutional layer in the encoder network are extracted and used to guide the high level features in the decoder part. Besides, all the parameters in this proposed mechanism are trainable and the model is trained from scratch.

Table 3
Proposed multispectral encoder configuration.

Block	Layer type	Kernel size	Number of channels	Output shape
	Input	–	–	W × H × 7
1	Convolution	3 × 3	64	W × H × 64
	Maxpooling	2 × 2		W/2 × H/2 × 64
2	Convolution	3 × 3	128	W/2 × H/2 × 128
	Maxpooling	2 × 2		W/4 × H/4 × 128
3 (Block 4 from FgSegNet_v2)	Convolution	3 × 3	512	W/4 × H/4 × 512
	Dropout			W/4 × H/4 × 512
	Convolution	3 × 3	512	W/4 × H/4 × 512
	Dropout			W/4 × H/4 × 512
	Convolution	3 × 3	512	W/4 × H/4 × 512
	Dropout			W/4 × H/4 × 512

What is needed to be stated is that, the filter channel of the first convolutional layer is not fixed in the proposed encoder, and will only be assigned when the images for training arrive. That is to say, for multispectral case, the size of the filter will be $3 \times 3 \times 7$ (7 stands for the number of available channels), while the size is set to $3 \times 3 \times 3$ if RGB images are fed. This property allows us to conduct a fair comparison for performance in background subtraction between multispectral images based model and RGB images based model.

4. Experiments

4.1. Evaluation dataset

To evaluate the performance of the proposed approaches for background subtraction, the multispectral dataset presented by Benezeth et al. [37] is adopted for testing in this paper. This dataset was created in order to investigate the use of multispectral videos of more than three bands for background subtraction.

To the best of our knowledge, this dataset is the only public real multispectral image background subtraction dataset available. Most public image datasets built for background subtraction, or change detection, such as the well-known Wallflower dataset [38], the Stuttgart Artificial Background Subtraction (SABS) dataset [39] and CDnet [34,35], are based on visible spectral images. Some other datasets include still recombined images. For example, the Grayscale-Thermal Foreground Detection (GTFD) dataset [40] provides a pair of grayscale and thermal frames captured by two cameras to investigate the fusion methods of thermal and grayscale data for effective foreground detection. Besides, the LITIV 2018 dataset [69] includes a pair composed of visible and Long Wavelength Infrared (LWIR) spectra.

As Benezeth et al. have not given an official name for the multispectral dataset they have established [37], some other works that use this dataset like [60,68] call it MSVS as an abbreviation of MultiSpectral Video Sequences. However, in [4], which is a survey work of datasets for background subtraction, this multispectral dataset is introduced as FluxData FD-1665 dataset, indicating the camera type for image acquisition. In this work, we will follow the name in [4], with a thought that more multispectral datasets will be built and published in future with a variety of multispectral imaging devices.

The FluxData FD-1665 dataset contains multispectral sequences with seven channels, or bands, captured simultaneously with the commercial camera from FluxData, Inc. (FD-1665-MS camera). Among the seven channels, six are in the visible spectrum and the last one is in the Near-InfraRed (NIR) spectrum. Fig. 6 shows an example associated with a single frame of one sequence of the multispectral dataset. In the

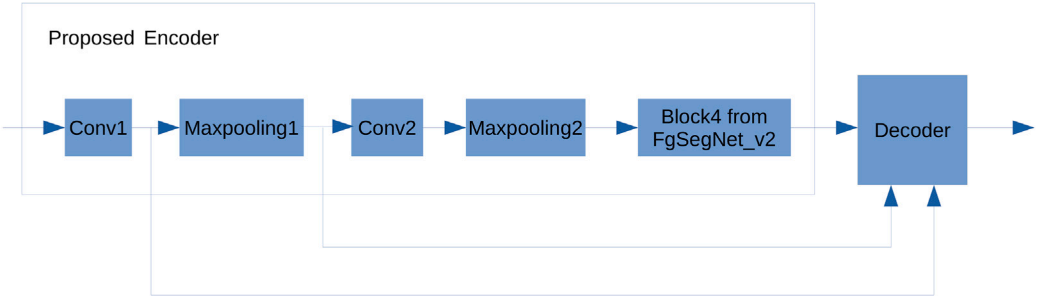

Fig. 5. Proposed multispectral background subtraction architecture.

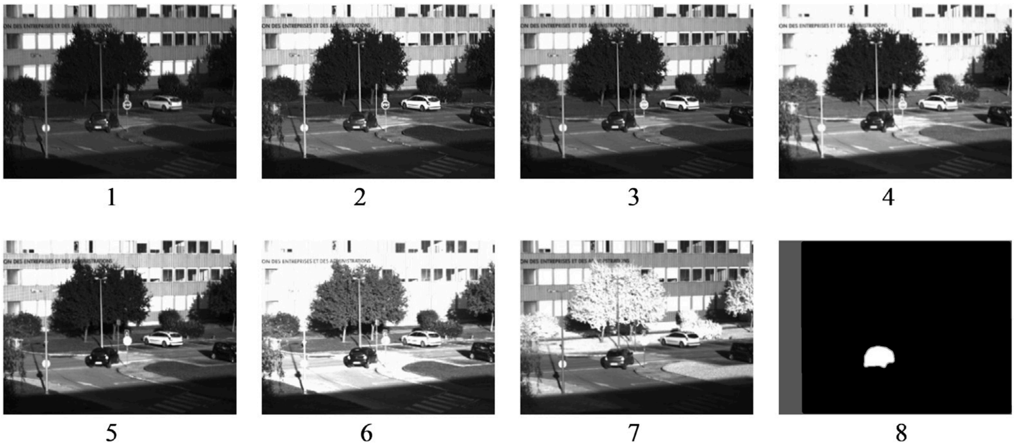

Fig. 6. Frame of one sequence of the multispectral dataset.

Table 4
Details of each video sequence of the FluxData FD-1665 dataset.

Video	Scene	Frames	Ground truth	Challenges
1	Indoor	258	258	Color saturation
2	Outdoor	1658	1658	Presence of shadows and camouflage
3	Outdoor	2213	1103	Dynamic backgrounds, shadows and occlusion
4	Outdoor	1351	925	Faraway intermittent object motion
5	Outdoor	3902	2232	Objects with shadows

figure, Images 1–6 show the six visible channels, Image 7 corresponds to the NIR channel, and the last image is the corresponding pixel-based ground truth.

The dataset is composed of five sequences containing frames of size 658 × 492 for each channel. The dataset represents one indoor video sequence and four outdoor scenes with different challenges such as shadows, intermittent object motion and camouflage effects (color similarity between object and background). The complete details of each video sequence are mentioned in Table 4.

The dataset includes as well the corresponding RGB sequences, which are obtained with a linear integration of the seven-channel original multispectral images weighted by three different spectral envelopes [70], shown in Eq. (9).

$$R=\sum_{i=1}^{n} r_i X_i,\ G=\sum_{i=1}^{n} g_i X_i,\ B=\sum_{i=1}^{n} b_i X_i \tag{9}$$

where r_i, g_i, b_i are the weights on the ith spectral channel, defined based on the characteristics of the specific multispectral camera, and n is the number of channels, which is seven here.

Fig. 7 presents examples of RGB frames extracted from the five RGB videos estimated from the five multispectral videos. Thus there are three subsets for each scene, namely multispectral image sequence with the size of 658 × 492 × 7, corresponding RGB image sequence of 658 × 492 × 3 and the groundtruth images of 658 × 492 × 1. These sequences are all publicly available.

Table 5
Ground truth labels.

Label	Grayscale value	Motion status
Moving	255	Foreground
Unknown	170	Half-occluded and corruptedby motion blur
Non-ROI	85	Unrelated activities
Static	0	Background

Pixel-wise labeled foreground masks, namely groundtruth images, are annotated manually and provided to public for a fair precise validation. With reference to [34], the groundtruth image is not an ideal binary mask, containing only moving pixels with a grayscale value of 1 and static pixels with a grayscale value of 0. Since some areas carry a certain level of uncertainty, it is difficult to reliably classify them as background or foreground pixels, such as those pixels close to moving objects boundaries, as illustrated with the partial magnification in the red circle in Fig. 8. To avoid evaluation metrics from being corrupted, these pixels that are corrupted by motion blur, are labeled as unknown and assigned grayscale value of 170.

Besides, the Non-ROI (not in the Region Of Interest) label is adopted to prevent the evaluation metrics being influenced by activities unrelated to the task considered. For example, the scene 3 is cluttered with moving tree, as shown in Fig. 8. However, what we care is the performance of an algorithm to detect the moving objects on the street. Thus the top and down parts are labeled as Non-ROI with a grayscale value of 85. The four labels are listed in Table 5, where pixels with a grayscale value of 85 and 170 in the ground truth images are ignored in accuracy evaluation.

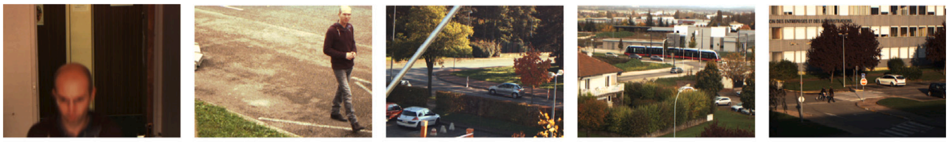
(1) Video 1   (2) Video 2   (3) Video 3   (4) Video 4   (5) Video 5

Fig. 7. Snapshots of RGB images extracted from the five RGB videos built from the five multispectral videos.

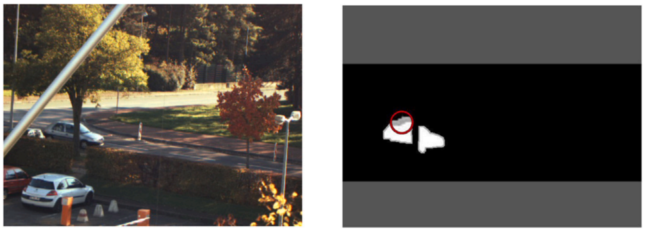

Fig. 8. Example of groundtruth labels.

4.2. Multispectral three-channel based results

The first experiments are conducted for thirty-five three-channel based combinations on the five different videos in the FluxData FD-1665 dataset. Given the groundtruth data, the performance of foreground detection is evaluated at a pixel level by F-measure, which is defined as a harmonic value of precision and recall with Eq. (10), where the relative contribution of precision and recall to the F-measure are equal. It is also known as balanced F-score or F1 score, which reaches its best value at 1 and worst score at 0. That is to say, the greater the value is, the better the detection quality is. It makes the analysis of results easier and is widely used in the domain of background subtraction or to compare different algorithms against a common dataset. What is needed to be noted is that only the pixels with the label of moving or static are taken into consideration in the evaluation process.

$$F-measure = 2\times\frac{Precision\times Recall}{Precision\;+\;Recall} \tag{10}$$

Table 6 shows the three-channel based F-measures on five videos, together with the RGB results in the last row, acting as a reference. The highest accuracy for each scene is in bold. It is clear that the best three-channel combination on one video is not always the same on the others, which is reasonable, since the spectral property and the ability to represent scenes vary with different channel combinations. This is quite obvious for Video 1, which is the only indoor scene, where the largest F-measure is 0.9703 (with the combination of Channels 4, 5 and 7), while the smallest is 0.9425 (with Channels 2, 4 and 5). As it is known, Channel 7 corresponds to the NIR spectrum and it offers good complementary information in this scene.

In order to be more clear in comparing the performance of RGB and other three-channel combinations extracted from the original seven-channel multispectral images, the largest F-measures for each scene in Table 6 are selected, together with the RGB for five videos and listed in Table 7. As Table 7 indicates, there are always three-channel based combinations that have better spectral discrimination and outperform the RGB images with a certain degree, which is in consistence with the results we have obtained utilizing the traditional Codebook method adapted to multispectral images [71–74]. This is more remarkable for the Video 5, where there exists objects with shadows. With the combination of Channels 1, 4 and 7, a higher F-measure of 0.9840 is obtained compared to 0.9714 for RGB images. This is quite interesting, because in this part we have followed the same mechanism of FgSegNet_v2 for the finetuning strategy, that is, only the parameters in the forth block are reinitialized and finetuned by the new data, while the first three blocks are frozen and the parameters are adopted directly from VGG16 deep model. As VGG16 is pretrained with conventional RGB images, one can expect better segmentation results with RGB images than the extracted three-channel based multispectral images. We think that the combination of channels (1, 4 and 7) performed better than RGB thanks, once again, to the NIR complementary spectral property of Channel 7.

Table 6
Three-channel based F-measures on five videos.

Channel combination	Video 1	Video 2	Video 3	Video 4	Video 5
123	0.9655	0.9983	0.9787	0.9666	0.9833
124	0.9618	0.9984	0.9804	0.9676	0.9799
125	0.9657	0.9983	0.9795	0.9688	0.9817
126	0.9667	0.9983	0.9785	0.9673	0.9830
127	0.9645	0.9983	0.9781	0.9695	0.9832
134	0.9675	0.9983	0.9790	0.9684	0.9828
135	0.9657	0.9983	0.9789	0.9693	0.9815
136	0.9675	0.9983	0.9781	0.9623	0.9832
137	0.9656	0.9983	0.9788	0.9669	0.9829
145	0.9629	0.9984	0.9780	0.9686	0.9828
146	0.9661	0.9982	0.9800	0.9659	0.9828
147	0.9653	0.9983	0.9783	0.9668	**0.9840**
156	0.9658	0.9983	0.9777	0.9645	0.9838
157	0.9638	0.9984	0.9790	0.9686	0.9818
167	0.9653	0.9984	0.9790	0.9682	0.9829
234	0.9480	0.9980	0.9818	0.9726	0.9765
235	0.9449	0.9981	0.9819	0.9706	0.9772
236	0.9445	0.9980	0.9809	0.9735	0.9795
237	0.9455	0.9982	**0.9835**	0.9729	0.9803
245	0.9425	0.9981	0.9820	0.9732	0.9800
246	0.9467	0.9982	0.9813	0.9715	0.9802
247	0.9451	0.9982	0.9817	0.9711	0.9818
256	0.9473	0.9979	0.9826	0.9709	0.9798
257	0.9455	0.9982	0.9813	0.9715	0.9753
267	0.9503	0.9982	0.9818	0.9757	0.9798
345	0.9584	**0.9986**	0.9829	0.9730	0.9759
346	0.9613	0.9986	0.9819	0.9731	0.9758
347	0.9611	0.9986	0.9833	0.9707	0.9787
356	0.9572	0.9985	0.9822	0.9730	0.9780
357	0.9624	0.9986	0.9809	0.9734	0.9758
367	0.9601	0.9986	0.9809	0.9735	0.9762
456	0.9669	0.9985	0.9797	0.9770	0.9786
457	**0.9703**	0.9984	0.9788	0.9774	0.9801
467	0.9692	0.9986	0.9797	**0.9788**	0.9785
567	0.9690	0.9986	0.9801	0.9754	0.9762
RGB	0.9683	0.9982	0.9808	0.9715	0.9714

Table 7
Best three-channel F-measures on five videos.

Video	Best MUL	RGB
1	**0.9703**	0.9683
2	**0.9986**	0.9982
3	**0.9835**	0.9808
4	**0.9788**	0.9715
5	**0.9840**	0.9714

4.3. Proposed convolutional encoder results

Subsequently, the experiments with the proposed convolutional encoder are conducted for both multispectral images and RGB images. As there is no pretrained deep model, all the parameters are trained from

Table 8
F-measures with the proposed convolutional encoder on five videos.

Video	MUL	RGB
1	**0.9786**	0.9540
2	**0.9982**	0.9942
3	**0.9430**	0.9109
4	**0.9619**	0.9558
5	**0.9603**	0.9383

Table 9
F-measures of different approaches on five videos.

Video	MD	SA	SID	OSTD	OWOC-RS	Superpixel-OWAOC	Proposed
1	0.8105	0.9042	0.9022	0.9365	0.9008	0.9135	**0.9786**
2	0.8900	0.9562	0.9686	0.9517	0.8727	0.9591	**0.9982**
3	0.6889	0.8970	0.8958	0.9064	**0.9635**	0.9376	0.9430
4	0.8327	0.6733	0.6878	0.8929	0.8997	0.8827	**0.9619**
5	0.7724	0.7422	0.7574	0.9266	0.8400	0.8693	**0.9603**
Mean	0.7989	0.8346	0.7427	0.9228	0.8953	0.9124	**0.9684**

MD = Mahalanobis Distance [37].
SA = Spectral Angle [37].
SID = Spectral Information Divergence [37].
OSTD = Online Stochastic Tensor Decomposition [60].
OWOC-RS = Online Weighted One-Class Random Subspace [67].
Superpixel-OWAOC = Superpixel-based Online Wagging One-Class Ensemble [68].

scratch. The only difference for multispectral images and the RGB case lies in the filter size of the first convolutional layer, namely, 3 × 3 × 7 for the former and 3 × 3 × 3 for the latter.

Table 8 illustrates the F-measures obtained with multispectral images based model and RGB images based model on the five videos in the FluxData FD-1665 dataset. As it is shown, multispectral images based model generally performs better than the RGB based one.

Specifically speaking, the proposed convolutional approach with multispectral images performs quite well on Video 2, where a very high F-measure can already be obtained by RGB images and there is no obvious accuracy difference between the two kinds of methods. However, for other videos, we could get considerable higher accuracy with multispectral images based model, especially for Video 3, where more than three percent improvement is obtained via the utilization of multispectral images, which is very impressive. The results show that multispectral images based model could be a promising alternative to conventional RGB images based one in background subtraction.

Besides, Fig. 9 shows some visual results. The top two rows are multispectral and RGB images, respectively. The third one is the corresponding groundtruth images. The forth and fifth rows are the background subtraction masks obtained by multispectral and RGB images, respectively. Since the pixels out of ROI or unknown are not considered in the training process, the detection results of these corresponding areas are random. In order to make the visual results tidy and easy to read, we assign the same gray value for these pixels in the mask images as they are originally in the groundtruth images.

We further compare the multispectral results obtained by the proposed convolutional encoder with the classical approaches in [37,60,67,68] using the same dataset. They have been explained in the former section. The F-measures of these different methods on the five videos of the FluxData FD-1665 dataset are collected and listed in Table 9.

The proposed convolutional approach outperforms the classical methods with a considerable gap on average, with a mean F-measure of 0.9684, which is nearly five percent higher than the ranking first classical algorithm OSTD in the fifth column. This shows the impressive ability of deep features learned with the proposed ConvNet in the task of background subtraction.

However, we need to be aware that the conventional feature selection methods can also obtain great accuracy with a carefully designed mechanism. As we can see, the OWOC-RS proposed by [67] has achieved the highest F-measure for the Video 3, where it exists dynamic backgrounds, shadows and occlusion. Since it is a challenging scene, we think the proposed multispectral images based model needs more images in the training set to better learn and represent the background. This supposal coincides with the fact that deep learning based methods always rely on big amount of data to achieve better performance. That is also part of the reason why there are still researchers devoted to classical methods or other new approaches, while deep leaning is changing the domain of computer vision nowadays.

4.4. Prediction time

Computational complexity is also observed during our experiments. The deep learning algorithm has been implemented using Keras framework with Tensorflow backend, on a single NVIDIA GeForce GTX1080Ti GPU with a memory of 11 GB and an Intel Core i7K-8700K CPU (6 cores and 12 threads) with a RAM of 32 GB. The time for prediction with our proposed approach is 0.134 s for each multispectral image with a resolution of 492 × 658. We have also tested our model on a i9-10900K CPU, the prediction time is 4.624 s per multispectral frame, which corresponds to a ratio of 34 when compared to the GPU prediction time (0.134 s).

To be best of our knowledge, our work is the first attempt to utilize multispectral images [37] for background subtraction task with deep learning based method. The authors of [60] have published their classical OSTD codes and provided the computational cost on specific resolutions with Matlab installed in laptop. For a more fair comparison, we have reimplemented this algorithm on our hardware operating on the same resolution as for our deep model, the prediction times for each frame are 1.237 s and 1.812 s on GPU and CPU, respectively. Comparing our method to the OSTD one when both are implemented on the same platform, we can see that our method is 9 times faster on GPU (0.134 s against 1.237 s) and 2.5 slower on CPU (4.624 s against 1.812 s). This conclusion is obvious, as neural networks based methods are more adequate thanks to their structure for GPU implementation than classical methods. On the other hand, they are slower when using CPU implementation.

5. Conclusion

In this work, we have followed the trend of deep learning and applied its concepts to background subtraction using multispectral images. Based on the ranking first algorithm, which is named FgSegNet_v2, on the large-scale change detection dataset CDnet, we have first extracted three channels out of the seven channels on five videos in the FluxData FD-1665 dataset to match the number of input channels for the pretrained VGG16 deep model with RGB images. The results are interesting, as some combinations of three-channel based multispectral images perform better than the conventional RGB images.

In order to further explore the benefits of multispectral images, we have proposed a new convolutional encoder for extracting the relevant deep features from the multispectral-groundtruth pair with images consisting any arbitrary number of channels. The modified VGG16 has been pretrained with by large-scale RGB ImageNet dataset, which are not available for multispectral case. Thus, there is no pretrained deep model adopted in the proposed encoder and all the parameters are trainable. The accuracy of the proposed convolutional approach is quite appealing when compared with other approaches using the same multispectral dataset. As the filter channel of the first convolutional layer is not fixed and can also be used for RGB images, the results show that the multispectral images based model outperforms the RGB images based one.

Our work can be seen as an attempt to investigate the potential advantages of using multispectral information via deep features learned with ConvNets for the background subtraction task. Some future research directions using multispectral images in background subtraction

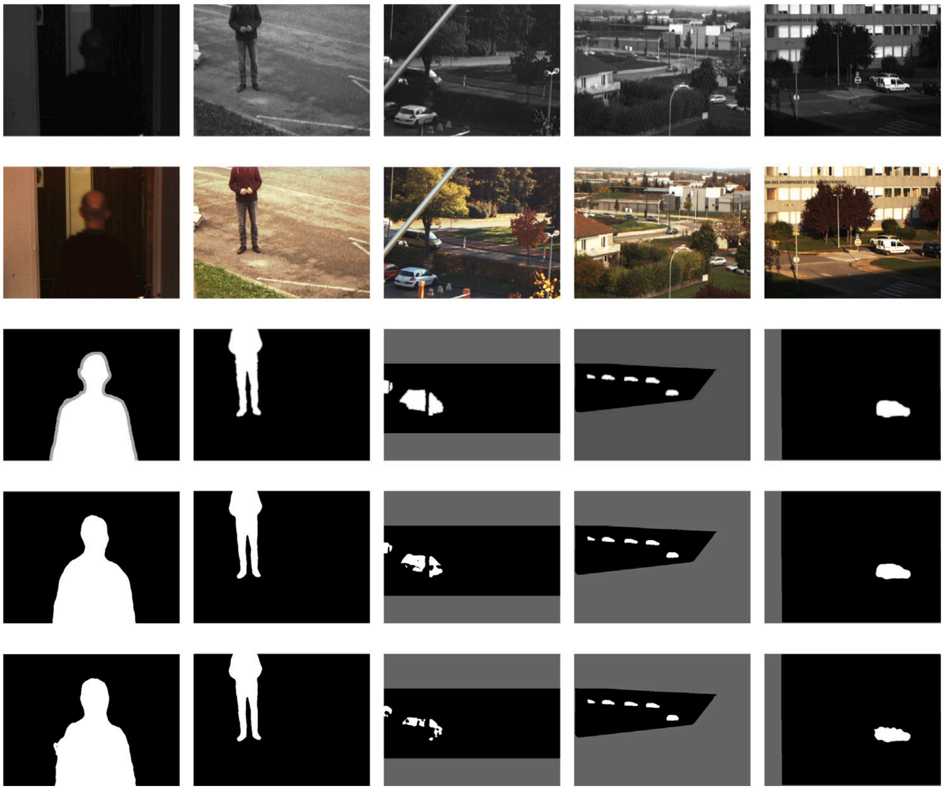

Fig. 9. Background subtraction results on the five videos.

are as follows. Larger multispectral datasets including other challenges, like night videos, should be investigated for exhaustive evaluation of multispectral information for background subtraction. Besides, as the filter channel of the first convolutional layer is arbitrary in the proposed encoder, the performance of background subtraction with other sizes (4, 5, 6) of multispectral images combination can be further studied. Another interesting direction lies in the combination of the proposed encoder and the pretrained deep models like VGG16 to take use of their powerful generic feature encoding properties. Last but not least, other data sources like depth, could also be explored and exploited for future research as they can offer important complementary information for background subtraction task.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge the financial support of the China Scholarship Council (CSC) for granting a PhD scholarship to the first author. Many thanks to Mohamed Kas, Assistant Professor in CIAD Lab., for his help in CPU and GPU implementation for computation time evaluation.

References

[1] M. Babaee, D.T. Dinh, G. Rigoll, A deep convolutional neural network for video sequence background subtraction, Pattern Recognit. 76 (2018) 635–649.

[2] R. Krungkaew, W. Kusakunniran, Foreground segmentation in a video by using a novel dynamic codebook, in: 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), IEEE, 2016, pp. 1–6.

[3] Y. Xu, J. Dong, B. Zhang, D. Xu, Background modeling methods in video analysis: A review and comparative evaluation, CAAI Trans. Intell. Technol. 1 (1) (2016) 43–60.

[4] R. Kalsotra, S. Arora, A comprehensive survey of video datasets for background subtraction, IEEE Access 7 (2019) 59143–59171.

[5] T. Bouwmans, B. Garcia-Garcia, Background subtraction in real applications: Challenges, current models and future directions, 2019, arXiv preprint arXiv:1901.03577.

[6] T. Yu, J. Yang, W. Lu, Refinement of background-subtraction methods based on convolutional neural network features for dynamic background, Algorithms 12 (7) (2019) 128.

[7] T. Bouwmans, Traditional and recent approaches in background modeling for foreground detection: An overview, Comp. Sci. Rev. 11–12 (2014) 31–66.

[8] A. Sobral, A. Vacavant, A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos, Comput. Vis. Image Underst. 122 (2014) 4–21.

[9] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), 2, IEEE, 1999, pp. 246–252.

[10] A. Elgammal, D. Harwood, L. Davis, Non-parametric model for background subtraction, in: European Conference on Computer Vision, Springer, 2000, pp. 751–767.

[11] K. Kim, T.H. Chalidabhongse, D. Harwood, L. Davis, Real-time foreground-background segmentation using codebook model, Real-Time Imaging 11 (3) (2005) 172–185.

[12] O. Barnich, M. Van Droogenbroeck, ViBe: A universal background subtraction algorithm for video sequences, IEEE Trans. Image Process. 20 (6) (2010) 1709–1724.

[13] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.

[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[15] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[18] J.R. Uijlings, K.E. Van De Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2) (2013) 154–171.

[19] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[20] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[21] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[22] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[23] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[24] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.

[25] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.

[26] L.A. Lim, H. Keles, Foreground segmentation using a triplet convolutional neural network for multiscale feature encoding, Pattern Recognit. Lett. 112 (2018).

[27] L.A. Lim, H.Y. Keles, Learning multi-scale features for foreground segmentation, Pattern Anal. Appl. (2019) http://dx.doi.org/10.1007/s10044-019-00845-9.

[28] M. Braham, M. Van Droogenbroeck, Deep background subtraction with scene-specific convolutional neural networks, in: 2016 International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, 2016, pp. 1–4.

[29] K. Lim, W.-D. Jang, C.-S. Kim, Background subtraction using encoder-decoder structured convolutional neural network, in: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2017, pp. 1–6.

[30] W. Zheng, K. Wang, F.-Y. Wang, A novel background subtraction algorithm based on parallel vision and Bayesian GANs, Neurocomputing (2019).

[31] W. Zheng, K. Wang, F. Wang, Background subtraction algorithm based on Bayesian generative adversarial networks, Acta Automat. Sinica 44 (5) (2018) 878–890.

[32] Y. Wang, Z. Luo, P.-M. Jodoin, Interactive deep learning method for segmenting moving objects, Pattern Recognit. Lett. 96 (2017) 66–75.

[33] M.C. Bakkay, H.A. Rashwan, H. Salmane, L. Khoudour, D. Puigtt, Y. Ruichek, BSCGAN: deep background subtraction with conditional generative adversarial networks, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 4018–4022.

[34] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, P. Ishwar, Changedetection. net: A new change detection benchmark dataset, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2012, pp. 1–8.

[35] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, P. Ishwar, CDnet 2014: an expanded change detection benchmark dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 387–394.

[36] A. Zheng, N. Ye, C. Li, X. Wang, J. Tang, Multi-modal foreground detection via inter-and intra-modality-consistent low-rank separation, Neurocomputing (2019).

[37] Y. Benezeth, D. Sidibé, J.-B. Thomas, Background subtraction with multispectral video sequences, in: IEEE International Conference on Robotics and Automation Workshop on Non-Classical Cameras, Camera Networks and Omnidirectional Vision (OMNIVIS), 2014, pp. 6–p.

[38] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, Wallflower: Principles and practice of background maintenance, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, 1, IEEE, 1999, pp. 255–261.

[39] S. Brutzer, B. Höferlin, G. Heidemann, Evaluation of background subtraction techniques for video surveillance, in: CVPR 2011, IEEE, 2011, pp. 1937–1944.

[40] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, L. Lin, Weighted low-rank decomposition for robust grayscale-thermal foreground detection, IEEE Trans. Circuits Syst. Video Technol. 27 (4) (2016) 725–738.

[41] W. Pitts, W.S. McCulloch, How we know universals the perception of auditory and visual forms, Bull. Math. Biophys. 9 (3) (1947) 127–147.

[42] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, Technical Report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[43] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition?, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 2146–2153.

[44] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.

[45] D.C. Ciresan, U. Meier, J. Masci, L.M. Gambardella, J. Schmidhuber, Flexible, high performance convolutional neural networks for image classification, in: Twenty-Second International Joint Conference on Artificial Intelligence, 2011.

[46] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, 2012, arXiv preprint arXiv:1207.0580.

[47] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015, arXiv preprint arXiv:1502.03167.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[49] Z.-Q. Zhao, P. Zheng, S.-t. Xu, X. Wu, Object detection with deep learning: A review, IEEE Trans. Neural Netw. Learn. Syst. (2019).

[50] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[51] M. Lin, Q. Chen, S. Yan, Network in network, 2013, arXiv preprint arXiv:1312.4400.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[53] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[54] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[55] A. Zhang, Z.C. Lipton, M. Li, A.J. Smola, Dive Into Deep Learning, 2019, http://www.d2l.ai.

[56] J. Dai, K. He, J. Sun, Instance-aware semantic segmentation via multi-task network cascades, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3150–3158.

[57] M. Ranzato, F.J. Huang, Y.-L. Boureau, Y. LeCun, Unsupervised learning of invariant feature hierarchies with applications to object recognition, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.

[58] R.A. Schowengerdt, Remote Sensing: Models and Methods for Image Processing, Elsevier, 2006.

[59] C.-I. Chang, An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis, IEEE Trans. Inform. Theory 46 (5) (2000) 1927–1932.

[60] A. Sobral, S. Javed, S. Ki Jung, T. Bouwmans, E.-h. Zahzah, Online stochastic tensor decomposition for background subtraction in multispectral video sequences, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 106–113.

[61] T. Bouwmans, A. Sobral, S. Javed, S.K. Jung, E.-H. Zahzah, Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset, Comp. Sci. Rev. 23 (2017) 1–71.

[62] M.A. Davenport, J. Romberg, An overview of low-rank matrix recovery from incomplete observations, IEEE J. Sel. Top. Signal Process. 10 (4) (2016) 608–622.

[63] Z. Lin, A review on low-rank models in data analysis, Big Data Inform. Anal. 1 (2/3) (2016) 139–161.

[64] J. Feng, H. Xu, S. Yan, Online robust PCA via stochastic optimization, in: Advances in Neural Information Processing Systems, 2013, pp. 404–412.

[65] J. Goes, T. Zhang, R. Arora, G. Lerman, Robust stochastic principal component analysis, in: Artificial Intelligence and Statistics, 2014, pp. 266–274.

[66] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, A review of microarray datasets and applied feature selection methods, Inform. Sci. 282 (2014) 111–135.

[67] C. Silva, T. Bouwmans, C. Frélicot, Online weighted one-class ensemble for feature selection in background/foreground separation, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2216–2221.

[68] C. Silva, T. Bouwmans, C. Frelicot, Superpixel-based online wagging one-class ensemble for feature selection in foreground/background separation, Pattern Recognit. Lett. 100 (2017) 144–151.

[69] P.-L. St-Charles, G.-A. Bilodeau, R. Bergevin, Online mutual foreground segmentation for multispectral stereo videos, Int. J. Comput. Vis. (2019) 1–19.

[70] N.P. Jacobson, M.R. Gupta, Design goals and solutions for display of hyperspectral images, IEEE Trans. Geosci. Remote Sens. 43 (11) (2005) 2684–2692.

[71] R. Liu, Y. Ruichek, M. El Bagdouri, Multispectral dynamic codebook and fusion strategy for moving objects detection, in: International Conference on Image and Signal Processing, Springer, 2020, pp. 35–43.

[72] R. Liu, Y. Ruichek, M. El Bagdouri, Extended codebook with multispectral sequences for background subtraction, Sensors 19 (3) (2019) 703.

[73] R. Liu, Y. Ruichek, M. El Bagdouri, Enhanced codebook model and fusion for object detection with multispectral images, in: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, 2018, pp. 225–232.

[74] R. Liu, Y. Ruichek, M. El Bagdouri, Background subtraction with multispectral images using codebook algorithm, in: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, 2017, pp. 581–590.