Deep Neural Networks for Moving Object Classification in Video Surveillance Applications

Rania Rebai Boukhriss[a,*], Emna Fendri[b], Mohamed Hammami[b]
[a]MIRACL-ISIMS, Sfax University, Sakiet Ezzit, Sfax, Tunisia
[b]MIRACL-FS, Sfax University, Road Sokra, Sfax, Tunisia

ABSTRACT

The moving object classification is a crucial step for several video surveillance applications whatever in the visible or thermal spectra. It still remains an active field of research considering the diversity of challenges related to this topic mainly in the context of an outdoor scene. In order to overcome several intricate situations, many moving objects classification methods have been proposed in the literature. Particular interest is given to the classes “Pedestrian” and “Vehicle”. In this paper, we have proposed a moving object classification approach based on deep learning methods from visible and infrared spectra. Three series of experiments carried on the challenging dataset “CD.net 2014” have proved that the proposed method reach accurate moving objects classification results when compared to methods based on deep learning and handcrafted features.

Keywords: Moving objects classification; Machine learning; Video surveillance; convolutional Neural Network.

1. INTRODUCTION

Moving object classification is a key step in intelligent video surveillance systems which consists in identifying the classes of detected objects in order to analyze their behavior. It remains an active research topic for several video surveillance applications such as abnormal events detection [1], people re-identification [2], human action recognition [3], vehicle license plate recognition [4], etc. The classification of moving objects in outdoor scenes is a challenging task, due to the complexity and diversity of real-world constraints. These challenges can be related to the environment, the moving objects or the acquisition equipment. Indeed, the variations in lighting conditions and dynamic weather conditions (*e.g.* overcast, fog, rain, snow) degrades the quality of images captured with visible cameras. In addition, the visibility of moving objects can be affected by the variation of the environment temperature in infrared images. Moreover, the moving objects present a set of intrinsic challenges including total or partial occultations, inter-object similarity and variation in appearance and/or resolution. In addition, the presence of shadow or halo, respectively, in visible or infrared image, modifies consequently the moving objects’ shapes. Finally, the change of point of view and the movement of the camera in a scene cause variation in the shape, speed, size and appearance of the moving objects and alter the classification accuracy. In the literature, several methods of classifying moving objects in the visible and infrared spectra have been proposed in the context of video surveillance in an outdoor scene. A particular interest is given to the classes “Pedestrian” and “Vehicle” who are the main classes used to analyze a scene. A study of existing works shows that the methods can be classified into two main categories: hand-crafted features-based methods and deep learning-based methods. On the one hand, the handcrafted features-based methods rely on convolutional descriptors. Thus, they can be categorized mainly into four main sub-categories: shape-based, texture-based, movement-based and hybrid methods, depending on the selected descriptor. On the other hand, the deep learning-based methods use generally a convolutional neural network to automatically extract deep features for moving object classification. In this paper, we have proposed a moving object classification method based on deep learning relying on infrared or visible spectrum. In our work, dealing with video surveillance, we have considered the two main classes of moving objects namely “Pedestrian” and “Vehicle”. The remainder of this paper is organized as follows: Section 2 provides the literature survey on moving object classification methods. Section 3 describes the proposed methods for moving object classification based on deep learning convolutional neural network model as a feature extractor. Experimental results of our work are outlined in Sect. 4. Finally, our conclusion and future works directions are introduced in Sect.5.

2. LITERATURE SURVEY OF MOVING OBJECT CLASSIFICATION METHODS

Moving object classification has received considerable attention in the computer vision research field given the variety of challenges related to this topic. Thus, several methods have been proposed to ensure accurate classification results. In this section, we provide a brief literature overview on moving object classification methods using whatever the visible or thermal spectra. According to previous research, existing methods for moving object classification can be classified in two main categories: handcrafted features-based methods and deep learning-based methods.

2.1 Handcrafted features-based methods

The handcraft features based methods have been developed for decades and still serve as a powerful tool when combined with machine learning classifiers [5] [6]. In fact, the robustness of these methods is related to the efficiency of the used descriptors. We can classify the most common methods into three major sub-categories based on shape, texture and / or movement. The shape-based methods use 2D spatial information such as area, bounding box size [7], silhouette, Histogram of Oriented Gradients (HOG) [8] and / or contours, etc. These methods are known to be sensitive to the presence of shadows/halos that affect the shape of the moving objects. In addition, they may fail to discriminate specific moving object classes (e.g., pedestrian groups and vehicles). On the other hand, texture-based methods rely on the spatial variation of the pixel's intensities of moving regions. It is about extracting the relevant texture characteristics by the appropriate descriptors such as Local Binary Pattern (LBP) [9], Markov Random Field (MRF) [10], etc. The study of texture-based methods has shown their performance in terms of precision in the task of objects classifying. However, these methods require a significant computing time [11]. Regarding the movement-based methods, they consider that object's movement patterns as well as temporal characteristics such as periodicity, direction and speed of tracked objects are relevant enough to classify the moving objects [12]. For instance, the residual flow can be used to analyze the rigidity and periodicity of moving objects. Rigid objects are expected to exhibit low residual flow while a non-rigid moving object has a higher average residual flow and even displays a periodic component [13]. Thus, the residual flow of a person or pedestrian group is differentiated from that of another rigid moving object, such as vehicles [14]. However, movement-based methods fail when pedestrians have an unusual gait pattern and in far-field images where movement of the legs and arms cannot be clearly captured. The movements-based features are discriminatory for the classification task, but can lead to false results in the case of slow movement. Moreover, several works in the literature have simultaneously combined shape, texture and / or motion to enhance the classification accuracy. In [15], the authors have combined shape and movement features for a better description of moving objects to discriminate between three moving objects classes namely "Pedestrian", "Pedestrian Group" and "Vehicle". As for Moctezuma et al. [16], they have combined a shape-based descriptor which is the Oriented Gradient Histogram (HoG) with a texture-based descriptor that is the Gabor filter. This combination aims to classify moving regions in human or non-human classes within uncontrolled surveillance environments.

2.2 Deep learning-based methods

During these last few years, deep learning techniques have been rapidly developed and have shown improvement on results in various tasks. Particularly, the deep learning methods based on Convolutional Neural Networks (CNN) have been used in the field of object recognition and classification, who recorded high performance [17]. The performance of CNNs come from the fact that they are capable to, automatically, extract information and learn features from image [18]. For instance, K. Simonyan et al. [19] have proposed a deep learning model "ConvNet" for large scale image classification. As for R. Shima et al. [20], they have proposed a deep convolutional neural network for object classification using depth images based on spatial information which is acquired by Kinect. However, the main drawback of CNNs is that they require massive data of training examples. For this reason, recent works have widely resorted to the transfer learning in the classification field due to the difficulty of collect large annotated datasets [21] [22]. In machine learning, the transfer learning can be classified in two categories related to the using way of the pre-trained network. Indeed, the first one consists in conserving the initial pre-trained CNN and adapting the parameters on the novel training data [23]. Whereas, in the second one, the pre-trained CNN is used to extract a set of discriminative features then an efficient classifier is employed to the classification task [24]. Thus, several CNN architectures (e.g., VGG-Net [19], AlexNet [25], ResNet [26], DenseNet [27], GoogleNet [28], ShuffleNet [29] and MobileNetV2 [30]) have been successfully applied for image features extraction and classification tasks. In this same framework, the authors of [31] have proposed an object recognition method with pose estimation based on transfer learning using the pre-trained CaffeNet model. Likewise, in [32] the authors have introduced a new method based on structured matching on the fast OverFeat pre-trained model to improve the image classification.

2.3 Discussion

Overall, the handcraft features extraction-based methods are designed to specific domain with a small database. Furthermore, deep learning methods have had a great surge in their popularity over the last few years due to their exceptional performance. Nevertheless, the deep learning methods require a large dataset in the offline phase to achieve outperforming accuracy [5] [33]. Hence, in order to accelerate the learning process, many works have resorted to a pretrained neural network that initializes the CNN with pretrained parameters rather than randomly set ones. Based on this study, we proposed in this paper, a deep features-based method for moving object classification. Indeed, we opted to transfer learning due to the lack of large visible/infrared annotated datasets.

3. PROPOSED METHOD

In this paper, we suggested a deep learning-based method for moving object classification. In fact, the classification relies on a deep learning convolutional neural network model as a feature extractor. Figure 1 shows the process of our method which consists of two main phases: the object model generation and the classification process, which are depicted in the following sub-sections.

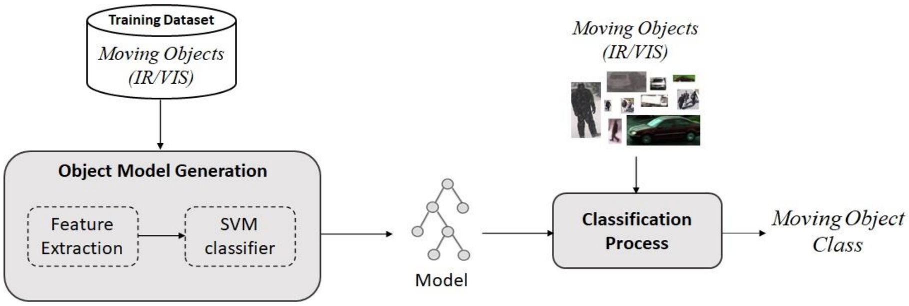

Figure 1. Proposed method for moving object classification.

3.1 Object model generation

The first phase is performed off-line. It aims to generate the prediction model for classifying moving objects into "Pedestrian" or "Vehicle" class. Hence, we start by a feature extraction step to extract the feature vector representing the moving objects who is fed, in a second step, to the SVM classifier in order to identify the class of each object.

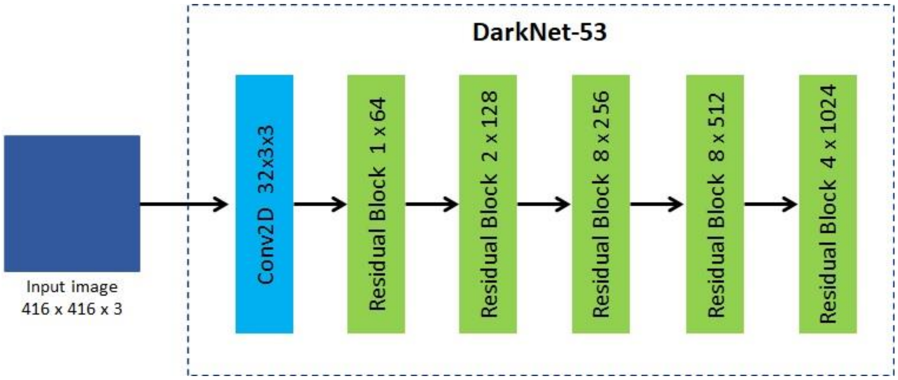

Figure 2. DarkNet-53 Architecture.

3.1.1 Feature extraction

As it is well known, the feature extraction is a crucial step in any classification problem that requires to identify discriminative features for training the classifier. Inspired by the high performance of convolutional neural network (CNN), we propose a moving object classification method based on deep features generated by a deep CNN model to select a powerful and discriminative features-set and an efficient representation of moving objects. In the literature, several CNN models have been proposed and trained on huge datasets. The Darknet-53 is a convolutional neural network which is part of the YOLOv3 as powerful feature extractor for object detection [34]. In fact, it was trained on ImageNet dataset and processes input image of size 416x416x3 pixels which is classified into the appropriate class. This architecture consists of 53 convolution layers which are composed by a layer's series at dimensions of 1×1 followed by 3×3. After each convolution layer, a batch normalization layer and a LeakyReLU layer is performed as well as a residual layer is introduced in order to overcome the problem related to the disappearance or the explosion of the network [35]. The object model generation consists firstly on training the network parameters to adapt the model to the moving object classification problem. Thus, we proposed to find-tune the DarkNet-53 CNN model (*cf.* Figure 2) to generate deep features and adapt it to our classification context.

3.1.2 SVM classifier

After finding the feature vectors of all training dataset, a classifier is fed to generate a prediction model allowing later the classes identification of each moving object. To this end, we have opted to an SVM classifier with linear kernel which is among the best classifiers designed for binary classification tasks.

3.2 Classification process

The goal of this online phase is to classify each detected moving object into the appropriate class. To extract the moving objects from a video stream whatever in infrared or visible spectrum, we resorted to our moving object detection method proposed in our previous research [36]. In fact, the proposed method detects moving objects based on background modelling in full-spectrum light sources (FSLS-MOD), which follows an appropriate temporal behavior. Indeed, our method finds its originality in the switching strategy between the infrared and visible spectra light sources according to the level of illumination and the state of the weather conditions. This method has recorded high accuracy proving the effectiveness and the feasibility of our system to run under different situations and various weather conditions. The classification is based on the model constructed in the object model generation step. In our work, we have considered two classes of moving objects namely “Pedestrian” and “Vehicle”, which are the main classes in the context of video surveillance in an outdoor scene. In order to demonstrate the performance of the proposed method several experiments were carried out and presented in the forthcoming section.

4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed method, we have resorted to three well-known datasets composed of challenging video sequences. The first set of data, named “OSU Color-Thermal Database” [37], consists of six pairs of infrared/visible video sequences which contains several pedestrians moving around the scene. The second corpus namely “INO Video Analytics Dataset” which is composed of ten infrared/visible video sequences covering different challenges [38]. The last dataset called “CD.net 2014” presents typical outdoor visible and thermal data containing large scope of challenges [39]. Relying on these challenging datasets, three series of experiments are conducted to assess the performance and the effectiveness of the proposed method for moving object classification. In fact, the first series of experiments validates the choice of the DarkNet-53 architecture as a discriminant feature extractor for moving objects classification in the visible/infrared spectra. While the second one, proves the effectiveness of the use of SVM classifier instead of the softmax classifier to classify the moving objects. Finally, in the third series of experiments, the classification performance is reported to prove the performance of deep features compared to handcrafted ones.

4.1 First series of experiments

The main purpose of this series of experiments was to validate the choice of the DarkNet-53 architecture as a discriminant feature extractor for moving objects classification. In fact, we have compared our classification accuracy recorded on the test dataset to other well-known deep CNN models which are widely used for the classification task in computer vision field (DarkNet-53, VGG-Net [19], AlexNet [25], ResNet-50 [26], DenseNet [27], GoogleNet [28], ShuffleNet [29] and MobileNetV2 [30]). These deep CNN models have been used for features extraction and the moving object classification was achieved with the same linear support vector machine classifier. Table 1 displays the classification accuracy.

Table 1. Comparison of CNN architectures for moving object classification in visible/infrared spectra in terms of accuracy.

CNN Architectures	Visible spectrum	Infrared Spectrum
VGG-19 [19]	97.26	80.71
AlexNet [25]	93.31	75.87
ResNet-50 [26]	97.68	75.73
DenseNet [27]	97.46	81.31
GoogleNet [28]	94.38	81.83
SuffleNet [29]	95.86	86.41
MobileNetV2 [30]	95.2	89.21
DarkNet-53 [35]	**98.12**	**91.54**

The presented results validate our choice of the DarkNet-53 deep architecture as feature extractor to provide an accurate representation of moving objects in both visible and thermal video stream. Note that the classification rate in thermal datasets is slightly weaker than those in visible ones due to the sensitivity of this spectrum to the environment temperature. Indeed, in the case of high temperature, the pedestrians can become somewhat invisible.

4.2 Second series of experiments

The aim of this experiment was to evaluate the performance of the use of linear SVM classifier in both the infrared spectrum and the visible one. Table 2 displays a comparison of the classification accuracy results obtained by the proposed method and those achieved by the softmax classifier.

Table 2. Comparison of classification accuracy obtained by the proposed method compared to those by the softmax classifier.

Methods	**Visible spectrum**	**Infrared Spectrum**
Softmax classifier	94.76	88.99
Proposed method	98.12	91.54

These reported results show that the use of linear SVM classifier have improved the results in both spectra. In fact, the classification accuracy has gone from 88.99 to 91.5 in the infrared spectrum and from 94.76 to 98.1 in visible spectrum. Such a statement shows the efficiency of using the linear SVM classifier together with deep features.

4.3 Third series of experiments

To confirm the effectiveness of our method, we compared our results to those obtained in our previous work [15]. In fact, in [15] we have proposed a moving object classification method based on handcrafted features which rely on shape and movement. Table 4 outlines the comparative study conducted on the “CD.net 2014” dataset. The linear SVM classifier was trained with these handcrafted features to select the appropriate class. The proposed deep learning-based method have shown a high performance in moving object classification as well in visible sequences as in thermal ones showing, once again, that the deep CNN model allows to generate discriminative features.

Table 3. Comparison of classification accuracy of the proposed method vs handcrafted features.

Methods	**Visible spectrum**	**Infrared Spectrum**
Handcrafted features [15]	91.84	72.68
Proposed method	98.12	91.54

5. CONCLUSION

The moving objects classification from visible or infrared video sequences stands a promising field. In this context, we have proposed a moving object classification method founded on a deep learning convolutional neural network model for feature extraction combined with a linear SVM classifier. Through the experimental evaluation of the proposed method on the “CD.net 2014” dataset, we have succeeded to validate its performance. Encouraged by the promising results, we aim to integrate the proposed method into a real-world video surveillance system. In addition, we will focus on other challenges mainly those related to the infrared spectrum.

REFERENCES

[1] T. Wang, Z. Miao, Y. Chen, Y. Zhou, G. Shan and H. Snoussi, “AED-Net: An Abnormal Event Detection Network,” Engineering Journal. **5**, 930-939, (2019)

[2] A.A. Sekh, D.P. Dogra, H. Choi et al., “Person Re-identification in Videos by Analyzing Spatio-temporal Tubes,” Mult. Tools Appl. **79**, 24537-24551, (2020)

[3] P. Pareek and A. Thakkar, “A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications,” Artificial Intelligence Review **54**, 2259-2322, (2021)

[4] R. Laroca et al., “A Robust Real-Time Automatic License Plate Recognition Based on the YOLO Detector,” International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 1-10, (2018)

[5] W. Lin, K. Hasenstab, G.M. Cunha et al., “Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment,” Sci Rep **10**, (2020)

[6] T. Kamnardsiri, W. Janchai, P. Khuwuthyakorn, P. Suwansrikham, J. Klaphajoneb and P. Suriyachan, "Knowledge-Based System Framework for Training Long Jump Athletes Using Action Recognition," **6**, 4, 182-193, (2015)

[7] S.B. Changalasetty, A.S. Badawy, W. Ghribi, H.I. Ashwi, A.M. Al-Shehri, A.D.A. Al-Shehri, L.S. Thota, and R. Medisetty, "Identification and classification of moving vehicles on road," Comp. Eng. and Int. Syst. **4**, 1-12, (2013)

[8] R. Tapu, B. Mocanu, A. Bursuc, and T. Zaharia, "A smartphone-based obstacle detection and classification system for assisting visually impaired people," IEEE Int. Conf. on Computer Vision Workshops, 444-451, (2013)

[9] F. M. Khellah, "Texture classification using dominant neighborhood structure," IEEE Trans. on Image Proc. **20**, 3270-3279, (2011)

[10] S. Yousefi and N. Kehtarnavaz, "A new stochastic image model based on markov random fields and its application to texture modeling," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1285-1288, (2011)

[11] H.S. Parekh, D.G. Thakore, and U.K. Jaliya, "A survey on object detection and tracking methods," Int. J. of Innovative Research in Computer and Communication Engineering **2**, 2970-2978, (2014)

[12] I. Shahu and R. Hablani, "An object detection and technique for velocity estimation along with classification on the basis of speed," Int. J. of Advanced Research in Computer Science and Software Engineering **6**, 233–236, (2016)

[13] H.A. Patel and D.G. Thakore, "Moving object tracking using kalman filter," Int. J. of Computer Science and Mobile Computing **2**, 326-332, (2013)

[14] P.K. Mishra and G. Saroha, "A study on classification for static and moving object in video surveillance system," Int. J. of Image, Graphics and Signal Processing **8**, 76-82, (2016)

[15] R.R. Boukhriss, E. Fendri and M. Hammami, "Moving object classification in infrared and visible spectra," Proc. SPIE 10341, 9th Int. Conf. on Machine Vision, 1034104, (2016)

[16] D. Moctezuma, C. Conde, I.M. de Diego, and E. Cabello, "Person detection in surveillance environment with hogg: Gabor filters and histogram of oriented gradient," IEEE Int. Conf. on Comput. Vis. Workshops, 1793-1800, (2011)

[17] S. Ahmed, MN. Huda, S. Rajbhandari, C.Saha, M. Elshaw, S. Kanarachos, "Pedestrian and Cyclist Detection and Intent Estimation for Autonomous Vehicles: A Survey", Applied Sciences **9**, 2335, (2019)

[18] D. Tome, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, S. Tubaro, "Deep Convolutional Neural Networks for pedestrian detection," Signal Process. Image Commun. **47**, 482-489, (2016)

[19] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv 2014, arXiv:1409.1556, (2014)

[20] R. Shima, H. Yunan, O. Fukuda, H. Okumura, K. Arai and N. Bu, "Object classification with deep convolutional neural network using spatial information," Int. Conf. on Intel. Inf. and Biomedical Sciences, Japan, 135–139, (2017)

[21] H. Azizpour, A.S. Razavian, J. Sullivan, A. Maki and S. Carlsson, "From generic to specific deep representations for visual recognition," IEEE Conf. on Comput. Vis. and Pattern Recognition Workshops, USA: 36-45, (2015)

[22] A.S. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," IEEE Conf. on Comput. Vis. and Pattern Recognition Workshops, USA: 512-519, (2014)

[23] S. Bunrit, N. Kerdprasop and K. Kerdprasop, "Improving the Representation of CNN Based Features by Autoencoder for a Task of Construction Material Image Classification," Journal of Advances in Information Technology, **11**, 4, 192-199, (2020)

[24] M.D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," Computer Vision, Lecture Notes in Computer Science, 8689, Springer, Cham, (2014)

[25] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Image net classification with deep convolutional neural networks," Adv. neural Inf. Process. Syst. **2012**, 1097-1105, (2012)

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In Proc. of the IEEE conf. on computer vision and pattern recognition, USA, 770-778, (2016)

[27] G. Huang, Z. Liu, L.V.D. Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA: 2261-2269, (2017)

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," IEEE Conf. on Computer Vision and Pattern Recognition, USA, 1-9, (2015)

[29] X. Zhang, X. Zhou, M. Lin and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," IEEE/CVF Conf. on Computer Vision and Pattern Recognition, USA, 6848-6856, (2018)

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," IEEE/CVF Conference on Computer Vision and Pattern Recognition, USA, 4510-4520, (2018)

[31] M. Schwarz, H. Schulz and S. Behnke, "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features," IEEE Int. Conf. on Robotics and Automation, USA, 1329-1335, (2015)

[32] C. Barat, C. Ducottet, "String representations and distances in deep Convolutional Neural Networks for image classification," Pattern Recognition, **54**, 104-115, (2016)

[33] K.A. Nugroho, "A Comparison of Handcrafted and Deep Neural Network Feature Extraction for Classifying Optical Coherence Tomography (OCT) Images," 2nd Int. Conf. on Inf. and Computational Sciences, Indonesia, 1–6, (2018)

[34] R. Joseph and F. Ali, "YOLOv3: An Incremental Improvement," arxiv:1804.02767Comment: Tech Report, (2018)

[35] M. Haojie, L. Yalan, R. Yuhuan, Y. Jingxian, "Detection of Collapsed Buildings in Post-Earthquake Remote Sensing Images Based on the Improved YOLOv3," Remote Sensing, **12**, 44, (2020)

[36] R.R. Boukhriss, E. Fendri and M. Hammami, "Moving object detection under different weather conditions using full-spectrum light sources," Pattern Recognition Letters **129**, 205–212, (2020)

[37] J.W. Davis, V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery", Computer Vision and Image Understanding, **106**, 162-182, (2007)

[38] L. St-Laurent, X. Maldague, D. Prevost (2007) "Combination of colour and thermal sensors for enhanced object detection", in 10th International Conference on Information Fusion: 1-8, (2007)

[39] Y. Wang, P. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, P. Ishwar, "Cdnet 2014: An expanded change detection benchmark dataset", IEEE Conference on Computer Vision and Pattern Recognition Workshops: 393-400, (2014)