End-to-end deep auto-encoder for segmenting a moving object with limited training data

Abdeldjalil Kebir[1], Mahmoud Taibi[2]
[1]Department of Electronics, Faculty of Sciences of engineers, Laboratory of Automatic and Signal Annaba,
University Badji Mokhtar, Annaba, Algeria
[2]Department of Electronics, Faculty of Sciences of engineers, Laboratory LERICA, University Badji Mokhtar, Annaba, Algeria

Article Info

Article history:

Received Sep 9, 2021
Revised Jun 5, 2022
Accepted Jul 2, 2022

Keywords:

Auto-encoder
Data augmentation
Limited data
Principal component analysis
Segmentation
VGG16

ABSTRACT

Deep learning-based approaches have been widely used in various applications, including segmentation and classification. However, a large amount of data is required to train such techniques. Indeed, in the surveillance video domain, there are few accessible data due to acquisition and experiment complexity. In this paper, we propose an end-to-end deep auto-encoder system for object segmenting from surveillance videos. Our main purpose is to enhance the process of distinguishing the foreground object when only limited data are available. To this end, we propose two approaches based on transfer learning and multi-depth auto-encoders to avoid over-fitting by combining classical data augmentation and principal component analysis (PCA) techniques to improve the quality of training data. Our approach achieves good results outperforming other popular models, which used the same principle of training with limited data. In addition, a detailed explanation of these techniques and some recommendations are provided. Our methodology constitutes a useful strategy for increasing samples in the deep learning domain and can be applied to improve segmentation accuracy. We believe that our strategy has a considerable interest in various applications such as medical and biological fields, especially in the early stages of experiments where there are few samples.

Corresponding Author:

Abdeldjalil Kebir
Department of Electronics, Faculty of Sciences of engineers, Laboratory of Automatic and Signal Annaba, University Badji Mokhtar
bp:12, 23000, Annaba, Algeria
Email: kebirabdeldjalil@gmail.com

1. INTRODUCTION

In the last years, deep learning architectures provided state-of-the-art results in various computer vision-related tasks, including image classification, object detection, and natural language processing (NLP) [1]–[3], to name a few. The deep learning concept is an artificial intelligence (AI) subfield which is different from machine learning techniques in how it learns representations from data. Unlike traditional machine learning techniques, deep learning models extract autonomously the hidden features from the data using a hierarchical network through numerous layers. Over the last few years, a wide range of deep learning architectures have been developed, examined, and discussed [3], [4]. In general, deep learning techniques may be divided into four main categories, namely: recurrent neural networks (RNNs), convolutional neural networks (CNNs), auto-encoders (AEs), and sparse coding [5].

Recently, deep learning models are becoming one of the most important concepts to solve several computer vision-related tasks, especially for image segmentation-based applications and dynamic background modeling [6]–[8], because it provides better performance over traditional machine learning methods. Several studies in the literature targeted the development of deep learning-based object segmentation models such as the works in [6], [9]. Most of these works studied the performance of deep learning segmentation methods in the case of using large datasets to train the deep learning model-based models. However, only a few studies explored how to train and enhance a deep learning model performance in the case of small datasets.

The segmentation of foreground regions that depict moving objects in videos is the core concept in most computer vision systems. Object segmentation is considered a crucial step, whereas, it presents a challenging task for many video surveillance applications like people counting, action recognition, and traffic monitoring [10]–[12]. Also, building an accurate model that is capable of segmenting moving objects in low-quality videos is even more challenging. In addition, other problems such as the presence of shadow, illumination change, dynamic background, and bad weather conditions can make the modeling task more complex. Moreover, the segmentation applied to small datasets remains a crucial challenge in computer vision which is also often the case in many real-world applications.

Training deep learning models on a small dataset has attracted particular attention in recent research studies in several fields. However, only a few works have addressed such a problem. For example, to overcome the problem of small dataset size, Salehinejad *et al.* [13] used a cylindrical transformation technique in a cylindrical coordinate system. Applying such transformations, they were able to make an object segmentation from 3D abdominal tomography achieving higher performance than the fully convolutional networks (FCNs) [14] in the case of using a limited number of annotated images. In order to mitigate the lack of training data, Keshari *et al.* [15] proposed an spectro-spatial feature-convolutional neural network (SSF-CNNN) architecture that modified the structure and strength of the filters obtained by CNN to reduce the number of learnable parameters. The proposed technique has proven its effectiveness for real-world newborn face recognition problems and multi-object classification. Salehinejad *et al.* [16] used a pixel-level radial transformation in a polar coordinate system for each image in order to increase the dataset samples' number. The proposed approach increased the models' generalization performance for various datasets.

The current study is part of a deep learning model developed for moving object segmentation. Deep learning is learned from data from the high-level features generated from the different network layers using simple learning methods. Furthermore, the presence of big databases is necessary in order to efficiently reconstruct the resulting segmentation mask and obtain better results from these precise features. However, in the real-world scenario, large databases are not always available. Based on this fact, we propose enhancing the precision of segmenting moving objects with little data training by using and comparing end-to-end auto-encoder through transfer learning and multi-depth techniques. Furthermore, most researchers use only traditional and general data augmentation techniques to enlarge the database such as (rotation, and translation) used in the general domain. Moreover, it is a common practice to fix the training data and change the model architecture. In the present work, some changes to the data are carried out to increase the number of samples. For this reason, we propose, compare and discuss object segmentation-oriented techniques to augment and enhance the quality of the training dataset that helps the model extract the relevant characteristics and cover the lack of necessary samples.

According to the fact that deep learning is a highly recommended topic in several fields, we conducted our work to be one of the first contributions that deals with the problem of training with little data in the area of deep learning. The results obtained from the comparative experiments between the proposed approaches and the well-known models show that the strategies used to improve and increase the database provide good results and help the model generalization. Hence, our work is considered to be an essential source of contribution to the research community and can be used in other areas when a large dataset is needed.

The rest of the paper is organized. In section 2, we present some theoretical basis for the used deep learning methods, concepts, and used materials. The proposed approaches and their performance evaluation are discussed in section 3. We discuss and conclude the paper in sections 4 and 5, respectively.

2. METHODS AND MATERIALS

In this section, we aim to present the methods and materials used to build a robust object segmentation system from video surveillance. We provide a detailed description of each of the main building blocks introduced in our methodology. In addition, the details of the proposed approach and the different data augmentation strategies are presented.

2.1. Auto-encoder

The auto-encoder is a successful deep neural network (DNN) type, which is considered among the unsupervised algorithms. It aims to reproduce the input data at the output [17], [18] where both the input and output layers have the same number of neurons. The auto-encoder consists of two main parts, which are the encoder and the decoder. The encoder's main role is to compress the input data into a lower-dimensional representation through the use of non-linear transformation while preserving the valuable features from the data by deleting the unnecessary elements. Then, the encoder outputs are fed to the decoder part to decompress them and reconstruct the original data from the generated lower dimension data. According to the literature, there exist four main auto-encoder architectures, including convolutional auto-encoder, variational auto-encoder, denoising auto-encoder, and sparse auto-encoder. Auto-encoders can be adopted in several applications like data denoising and dimensionality reduction [19]. Figure 1 shows the overall network architecture.

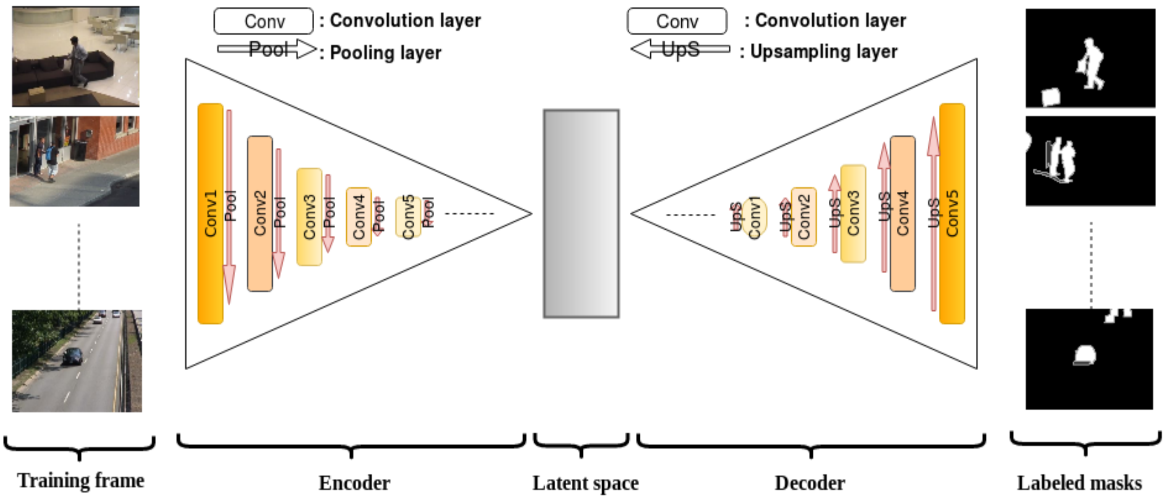


Figure 1. The architecture of the general auto-encoder approach

The auto-encoder has three essential blocks: i) encoder part: the encoder aims to encode all of the relevant information about the input in the latent space; ii) latent space: it represents the space represented by a compressed form of the input; and iii) decoder part: the decoder aims to reproduce the input data at the output level by focusing only on the data in the latent space. Encoding the input data X with nonlinear encoder function E to Z=E(X), then decoding z to Y=D(Z) through nonlinear decoder function D which approximates the original data X. As shown in Figure 2. We can describe this algorithm in its simplest form as (1).

$$Y = D(E(X)) \tag{1}$$

The learning process minimizes the loss function between the input X and output Y as (2).

$$Loss(\mathrm{X},\mathrm{Y}) = \left|\left|\mathrm{X}-\mathrm{Y}\right|\right|^2 = \left|\left|\mathrm{X}-\mathrm{D}(\mathrm{Z})\right|\right|^2 = \left|\left|\mathrm{X}-\mathrm{D}(\mathrm{E}(\mathrm{X}))\right|\right|^2 \tag{2}$$

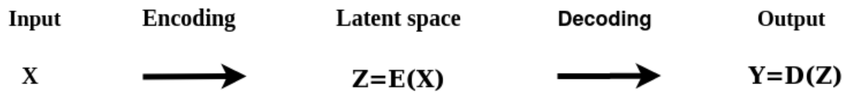


Figure 2. The basic expression of a general auto-encoder

2.2. Transfer learning

Transfer learning is an interesting approach to training efficient deep learning models when only small datasets are available. Compared to deep learning models trained from scratch, the transfer learning technique aims to improve the model accuracy using lower computational power. The transfer learning concept could be considered as a two-step technique. Firstly, it aims to learn data representation by training a model on available datasets containing a large number of annotated data. Then, it uses this representation to build a new model based on the pre-trained model using a smaller dataset, by training only some selected layers or the final decision layer [20], [21].

Transfer learning [22], [23] is a machine learning method, where a model developed for a given task is reused as a reference for another model on a second task. The concept is to use the knowledge learned from the first model when solving a new problem. In other words, we can say that it is a transfer of knowledge. However, the benefit of using transfer learning is that large dataset training is not needed to avoid over-fitting and not many computational resources are required.

2.3. Data augmentation

The performance of deep learning models depends on the size of the training dataset. However, the lack of available datasets, in several fields, is one of the most critical issues facing researchers. To overcome such a problem, several solutions have been proposed over the years, including transfer learning and data augmentation. To this end, in the current study, we tested dataset augmentation [20]. Data augmentation is a procedure that aims to enlarge the dataset size by applying some transformations, where both the original and the created images are used to train the model [24]. Therefore, our main objective is to use the existing dataset to generate new data to avoid the over-fitting problem while improving the model performance. One of the main data augmentation techniques is to perform some adjustments and geometric transformations, including cropping, translation, scaling, mirroring, rotating, and changing lighting conditions. These methods are widely used in the literature to solve problems related to image and video processing, including detection, recognition, and segmentation, to name a few.

2.4. Principal component analysis

Principal component analysis (PCA) [25] is an unsupervised technique based on simple linear transformation, it is a dimensionality reduction technique [26]. However, the main goal of a PCA is to compress data. It is used in many applications of image processing such as image compression [27] and face recognition [28]. Indeed, in our case, it will be used as new feature extraction compression and reconstruction technique to preserve and extract new and essential features linearly in various levels of the distribution of the data with the aim of using it as a new data augmentation technique.

The implementation used for the reconstruction and compression of color frames using PCA can be divided into 3 main steps: i) splitting the frames into 3 channels R, G, and B arrays; ii) performing the PCA and selecting the most dominated N eigenvalues on each color value matrix; and iii) recreating the original frames by merging the R, G, and B components.

2.5. Evaluation protocol

The proposed approaches in this study are based on transfer learning and multi-depth auto-encoder. To perform the segmentation of the moving object, we employ the auto-encoder as supervised learning for both approaches. For the first approach, we construct the network by fine-tuning the VGG-16 network [29], [30] that was pre-trained trained on the famous ImageNet dataset [29], [31] as the encoder part. Then, we changed the fully connected layers with a latent space. On the other hand, the transposed architecture of the VGG-16 has been used in the decoder part to reconstruct the resulting mask of the input frames. The reconstruction process aims to increase the encoder output size to reconstruct the original input data through upsampling and convolution operations, which are called transposed VGG-16 architecture. Finally, we only train the latent space and the decoder part with the CDnet2014 dataset [32] while there is no training process on the pre-trained encoder part.

In the second approach, convolution and pooling layers are stacked to build the encoder part, whereas, upsampling layers are used in the decoder part to up-sampling the images in the latent space. The hidden layers are in multi-depth. For this approach, we have trained the whole model with the CDnet2014 dataset. Figures 3 and 4 show the overall architectures proposed in the current study, which are based on transfer learning and multi-depth auto-encoder architectures.

2.6. Dataset and metrics

The Cdnet2014 dataset (change detection) [32] is adopted to train and test our model. It consists of real videos captured in challenging scenarios as shown in Table 1. For further generalization of the training process, we select all video sequences (53 scenes), which contain 11 video categories from the CDnet2014 dataset; each video has an average of 2,000 frames.

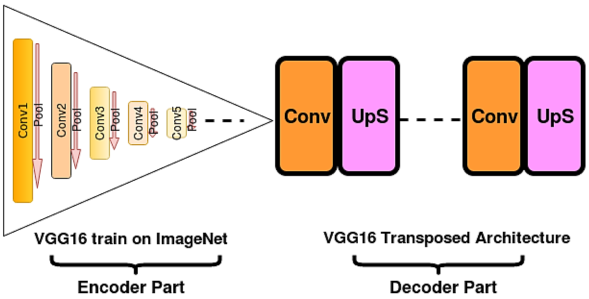


Figure 3. Structure layers of the transfer learning-based model approach

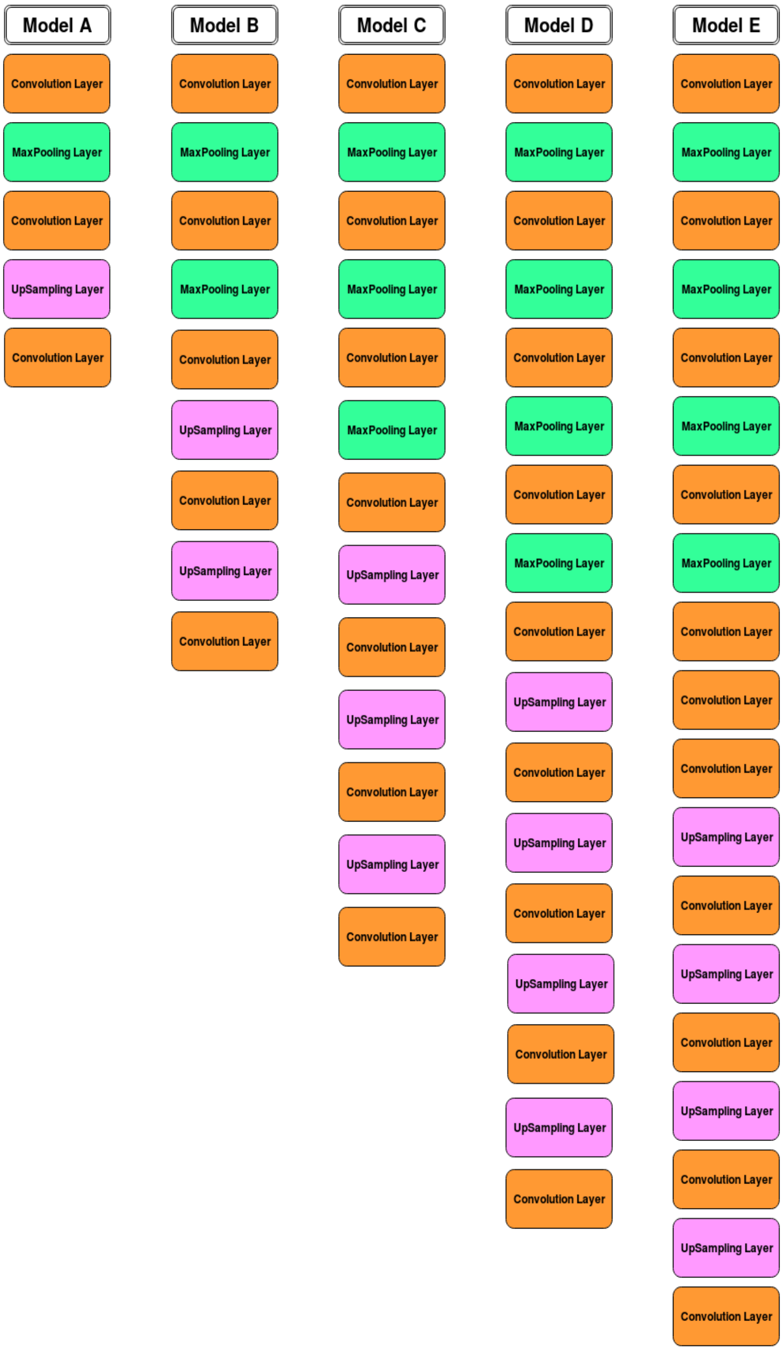


Figure 4. Detailed information of five models at different depths

Table 1. Shows the list of categories and video names in the CDnet 2014 dataset

Categories/Challenges	Video names
Baseline	Highway, Office, Pedestrians, PETS2006
Camera Jitter	Badminton, Sidewalk, Traffic, Boulevard
Bad Weather	Skating, Wet snow, Blizzard, Snowfall
Dynamic Background	Boats, Canoe, Fountain1, Fountain2, Fall, Overpass
Intermittent Object Motion	Abandoned box, Street light, Parking, Sofa, Tram stop, Winter driveway
Low Frame rate	Port_0_17 fps, Tram crossroad_1 fps, tunnel exit_0_35 fps, Turnpike_0_5 fps
Night Videos	Bridge entry, busy boulevard, fluid highway, Street corner at night, Tram station, Winter street
PTZ	Continuous pan, Intermittent pan, Two-position ptz cam, Zoom in zoom out
Shadow	Back door, Copy machine, Bungalows, Bus station, Cubicle, People in shade
Thermal	Corridor, Library, Lakeside, Dining room, Park
Turbulence	Turbulence0, Turbulence1, Turbulence2, Turbulence3

Several metrics are adopted to evaluate the deep learning-based models [32], including specificity, precision, f-measure, false positive rate, false negative rate, and percentage of wrong classifications, by using the four parameters of the confusion matrix. These metrics can be measured according to:

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{3}$$

$$\text{False Positive Rate} = \frac{FP}{TN+FP} \tag{4}$$

$$\text{False Negative Rate} = \frac{FN}{TP+FN} \tag{5}$$

$$\text{Percentage of the Wrong Classifications} = 100 \times \frac{FN+FP}{TP+FN+FP+TN} \tag{6}$$

$$\text{F} - \text{Measure} = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

where TP, FP, FN, and TN denote true positive, false positive, false negative, and true negative, respectively.

2.7. Training data process

In this sub-section, we present the training data process of our data selection strategies, we manually chose twenty-five frames for each video that contain important foreground objects to help our model learn and segment the foreground accurately. However, the model has been fed across a variety of data training strategies, including: i) the PCA strategy: this strategy consists of generating four different projections of the principal components (geometric transformation) using the PCA technique at a rate of 80%, 60%, 40%, and 20% to eliminate non-informative variables, for each selected training frame; ii) the data augmentation (DA) strategy: the DA strategy aims to enlarge the data by generating four morphological transformation frames for each selected training frame, including translating, flipping, zooming, and rotating; iii) the PCA and DA strategy: this strategy consists of merging the aforementioned strategies (PCA and DA) to build more data frames for each video. Figure 5 shows the results of various transformation techniques and strategies. Table 2 provides the description and the amount of samples used for the training process of all strategies.

In our experiments, we perform our implementation using an open-source library called Keras which was developed in 2018 by Chollet *et al.* [33]. The training process is done on the Google Colaboratory platform through a Tesla K80 GPU [34], [35] for 100 epochs. We selected the RMSprop as the main optimizer to train our model. Binary cross entropy (BCE) loss function is used to compute the loss between the ground truth label and the predicted result, which can be measured using (9):

$$BCE(Y, \widetilde{\mathrm{X}}) = -(Y * \log(\tilde{X}) + (1-Y) * \log(1-\widetilde{\mathrm{X}})) \tag{9}$$

where X̃ and Y denote the ground-truth label and the label predicted by the models, respectively. We train the networks with 80% frames from the training data and 20% frames as validation. We evaluate the models with 50% of the dataset.

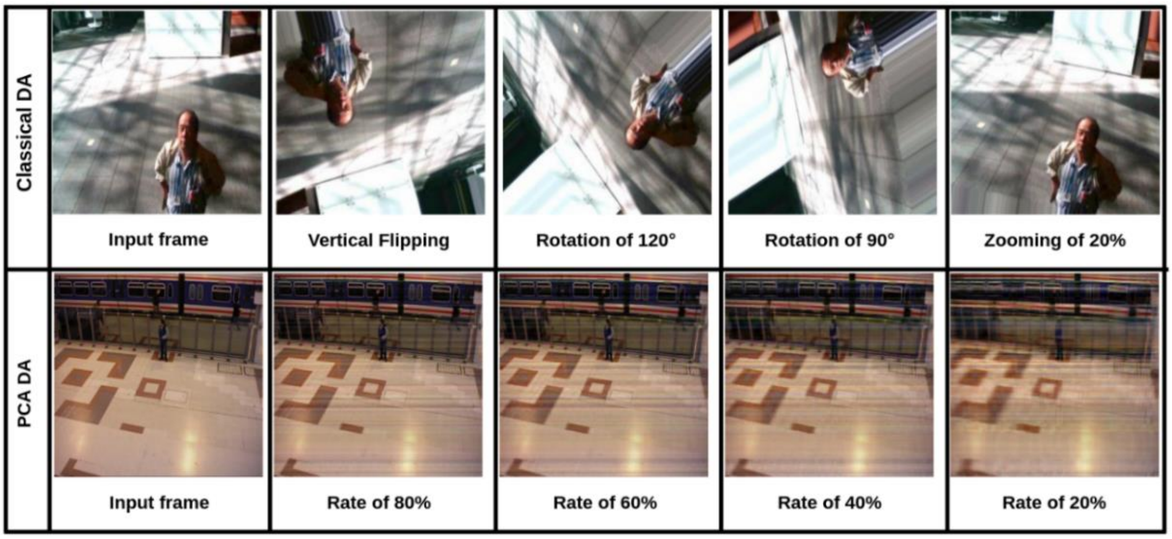


Figure 5. Examples of the different strategies transformation

Table 2. Description and number of training samples used for each strategy

Strategy	Description	Number of samples used to train the model
DA	25(frames) x 53(all video) x 5	6625
PCA	25(frames) x 53(all video) x 5	6625
PCA+DA	25(frames) x 53(all video) x 9	11925

3. EXPERIMENTAL RESULTS

In the current section, we aim to present the implementation and the achieved results using our proposed approaches. Moreover, to illustrate the effectiveness of our models, we compare them with the conventional algorithms. More detail is described in subsections 3.1 to 3.3.

3.1. Experiments

For the first approach, we freeze the first 14 layers of the VGG16 (encoder part), then we execute the training for the remaining layers of the latent space and all the decoder part (VGG16 transposed). The dropout layer [31] applied after every convolution layer of the decoder part is set to a learning rate of 0.2. Figure 3 shows an explanatory diagram.

We developed five models in the second approach, starting with four hidden layers and eventually increasing to eighteen hidden layers. The dropout layer [31] applied after every convolution layer is set to a learning rate of 0.2 to generalize the model. Figure 4 shows the detailed information and structure concerning the layers of the different multi-depth approach models.

3.2. Evaluation of the proposed approaches

In this sub-section, we analyze the training strategies using the obtained results and present their influences on the adopted dataset. Subsection 3.2.1 explained about transfer learning approach. Subsection 3.2.2 explained about multi-depth approach.

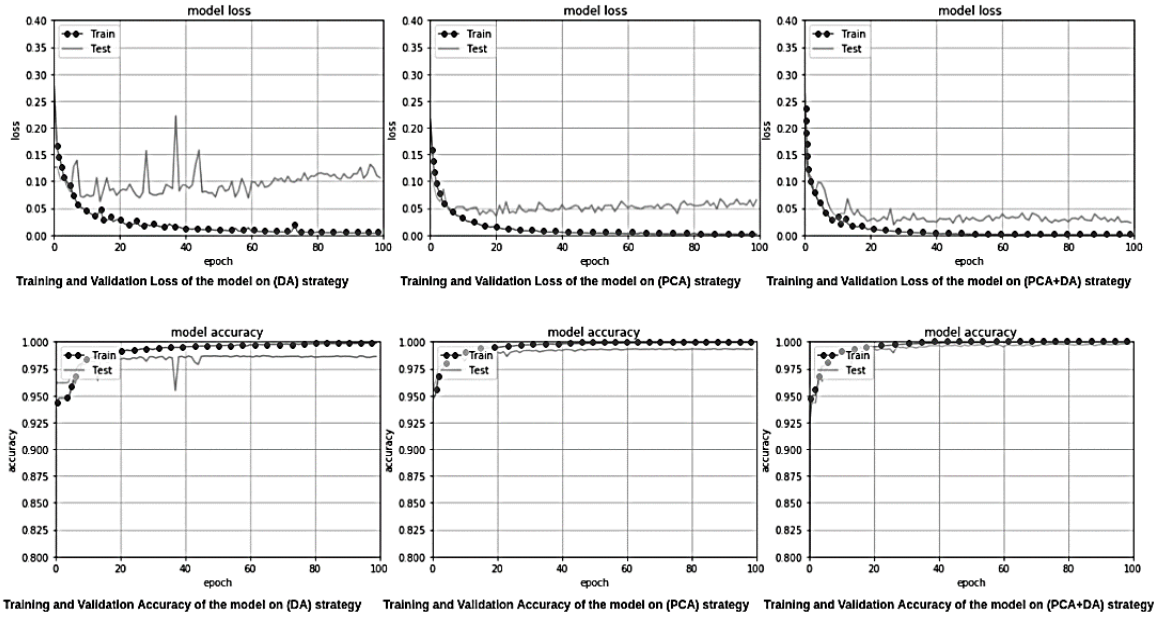
3.2.1. Transfer learning approach

The obtained results using the transfer learning approach are shown in Table 3 and Figure 6. The results clearly show that the PCA+DA strategy outperformed the other strategies providing better performance. Figure 6 shows the training and validation accuracies and losses graphs.

Table 3. The test results obtained by PCA, DA and PCA+DA strategies

Strategy	Sp	FPR	FNR	PWC	FM	Pr
PCA	0.993	0.0058	0.3458	1.383	0.692	0.748
DA	0.997	0.0029	**0.3338**	1.073	0.742	0.853
PCA+DA	**0.997**	**0.0020**	0.3571	**1.027**	**0.811**	**0.873**

According to the loss graph of the classical DA strategy as shown in Figure 6, we can clearly see the gap between validation and training loss indicating that the model is over-fitting. The over-fitting could be due to the lack of training samples. Also, as shown in Figure 6, the gap between train and test losses is reduced in the case of using the PCA strategy, and there is a slightly difference between training and validation loss values. Therefore, we can see that the over-fitting effect has been reduced due to the presence of PCA transformation with higher precision features which help and facilitate the system for the learning task.



Training and Validation Loss of the model on (DA) strategy

Training and Validation Loss of the model on (PCA) strategy

Training and Validation Loss of the model on (PCA+DA) strategy

Training and Validation Accuracy of the model on (DA) strategy

Training and Validation Accuracy of the model on (PCA) strategy

Training and Validation Accuracy of the model on (PCA+DA) strategy

Figure 6. Used strategies training and validation accuracies and losses graphs

The proposed PCA+DA strategy provides the best and lowest training and testing losses over the other two strategies as shown in Figure 6. Thus, the model avoids over-fitting through the number of samples added to the training data-set. For the accuracy curves in Figure 6, observe that with each increase of the PCA samples in the training data the accuracy increases and both curves converge even more.

3.2.2. Multi-depth approach

The current study sought to evaluate the auto-encoder model in various depths, we analyzed the influence and the results of the multi-depth training model. For that, both loss function and accuracy function curves between training and validation data have been plotted. We fed the five models only with the DA strategy. The results are shown in Figures 7 and 8.

We can see that for the models (A), (B), and (C), the validation and training losses are very close but have a higher error. Validation and training accuracy have a low precision value. According to Figure 8, we can see that the models (D) and (E) starts to over-fit, and the error begins to decrease. The validation accuracy and training accuracy start to increase.

Generally, when we increase the number of layers, it may result in better accuracy and minimum loss, increasing the depth means increasing the capacity of the model (can learn complex representations), but with little training data, it may cause a high risk of over-fitting. We show that the model (E) provides better results than the other methods in terms of accuracy and error. However, the model (E) starts to over-fit from around epoch 50. To this end, we selected the model (E) as the base model to be improved by adding more training data. The results of the improved version are shown in Figure 9.

The model (F) is the same model (E) but trained with more data. In addition to the data generated using the classical DA techniques, the data used to train the model (F) is generated using the PCA strategy. Hence, the model is trained using both DA and PCA strategies. As shown in Figure 9, we can notice that the validation loss and accuracy of the model (F) are improved making it more robust to over-fitting. Our main goal is to achieve high performance in terms of accuracy and loss using deep CNN architectures with small datasets while avoiding over-fitting.

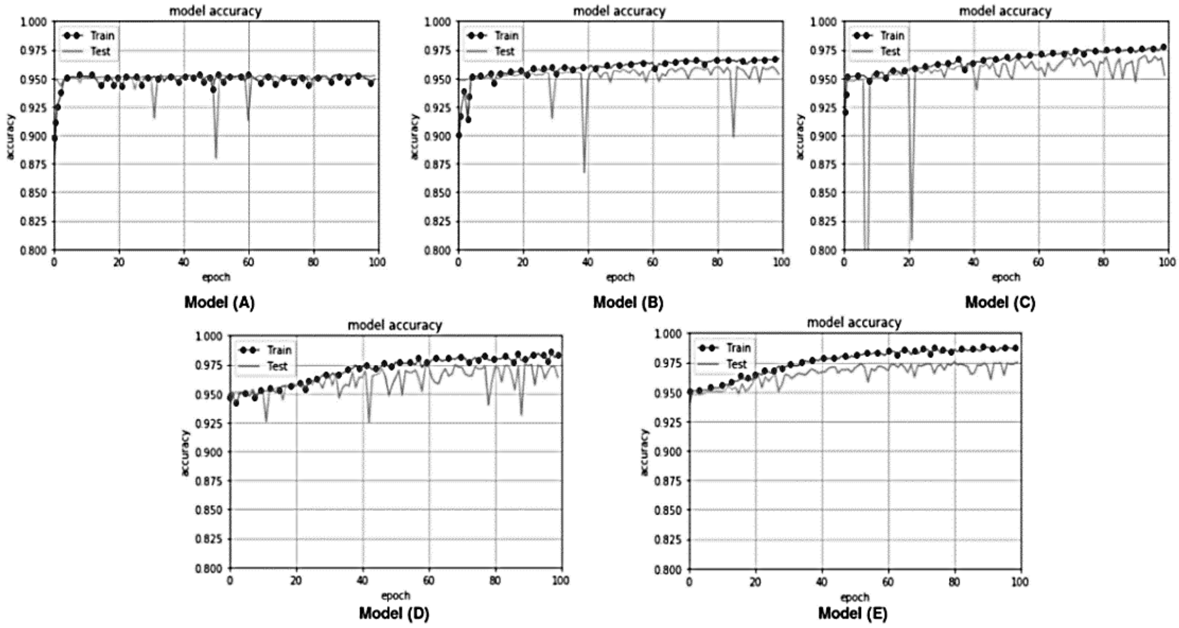


Figure 7. Used strategies training and validation accuracy graphs

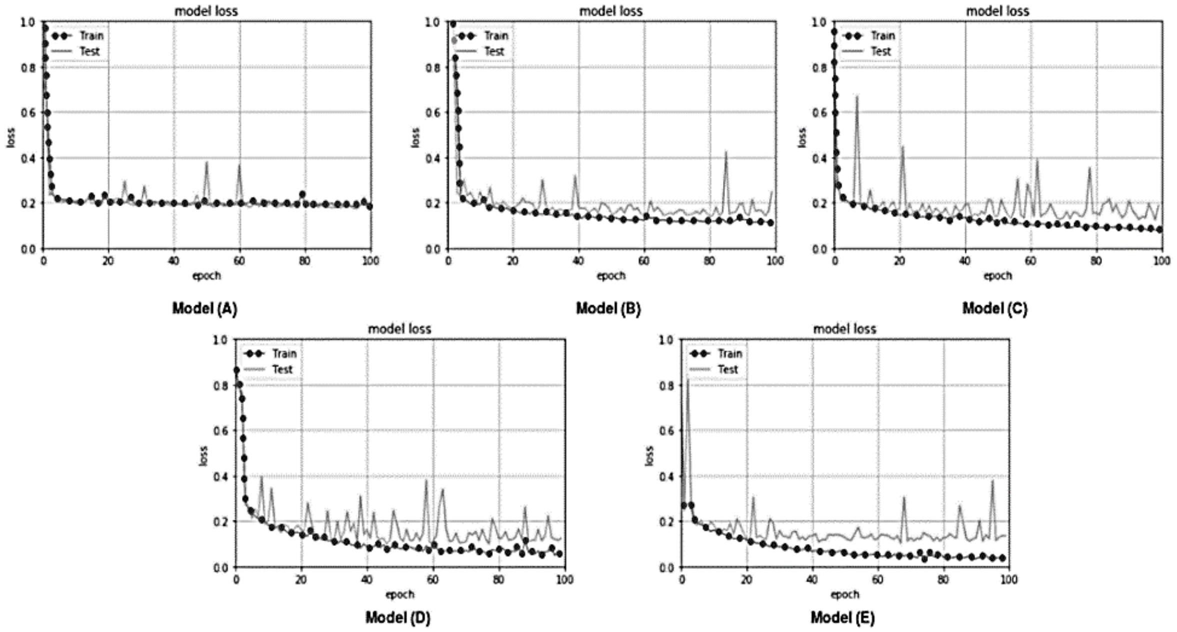


Figure 8. Used strategies training and validation loss graphs

3.3. Comparison with reference algorithms

The amount of training data that is used to produce the models are also different in different approaches. Since that, we compared our models with methods that use the same principle of training with few training data. We compared our model with the one developed by Babaee *et al.* [9]. This model is trained by 5% (100 frames) of frames from each video sequence. Furthermore, both BSUV-Net [36] and fast BSUV-Net 2.0 [37] proposed background subtraction algorithms for unseen videos based on a fully CNN. They introduced a spatio-temporal data augmentation technique to overcome the lack of training samples issue. Also, our approach is compared with other traditional algorithms, including SuBSENSE [38], IUTIS-5 [39], and pan-arctic water-carbon cycles (PAWCS) [40], where the models are trained through little frames from each video sequence of the dataset. We compare with our best models for each approach, both multi-depth auto-encoder (MD AE) and transfer learning auto-encoder (TL AE) training with the PCA+DA strategy. The results obtained using our models are compared with other studies as shown in Table 4.

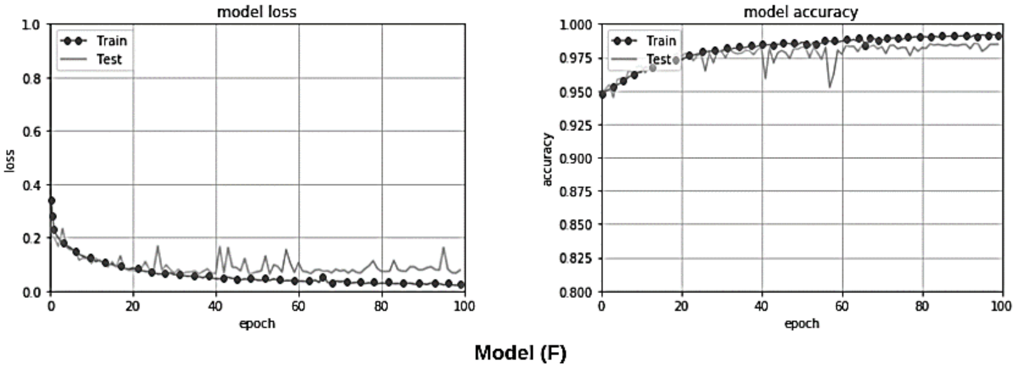


Figure 9. Used strategies training and validation accuracy and loss graphs

Table 4. A comparison of our result approach with reference algorithms

Method	Sp	FPR	FNR	PWC	FM	Pr
TL AE	**0.997**	**0.0020**	0.3571	1.027	**0.811**	**0.873**
Fast BSUV-Net 2.0 [37]	0.995	0.0044	0.1819	**0.905**	0.803	0.842
BSUV-Net [36]	0.995	0.0054	**0.1797**	1.140	0.787	0.811
CNN [9]	0.990	0.0095	0.2455	1.992	0.755	0.833
MD AE	0.990	0.0071	0.4467	1.836	0.747	0.809
IUTIS-5 [39]	0.995	0.0052	0.2151	1.198	0.772	0.808
PAWCS [40]	0.995	0.0051	0.2280	1.199	0.740	0.786
SuBSENSE [38]	0.990	0.0096	0.1876	1.678	0.741	0.751
TL AE	**0.997**	**0.0020**	0.3571	1.027	**0.811**	**0.873**

For further evaluation, we compare our approaches with other state-of-the-art models. Figure 10 provides qualitative comparison results. Three frames of video sequences from the CDnet 2014 dataset are selected as demonstrative examples. The first column and second columns in Figure 10, shows the input frames and the ground truth, respectively. The third column presents our deep Auto-encoder model based on transfer learning trained through (PCA+DA) strategy. Whereas, the rest columns in Figure 10 represent the results of the reference models. The results from Table 4 and Figure 10 show that our approach based on the PCA+DA training strategy provides better results and selectivity segmentation than the other models.

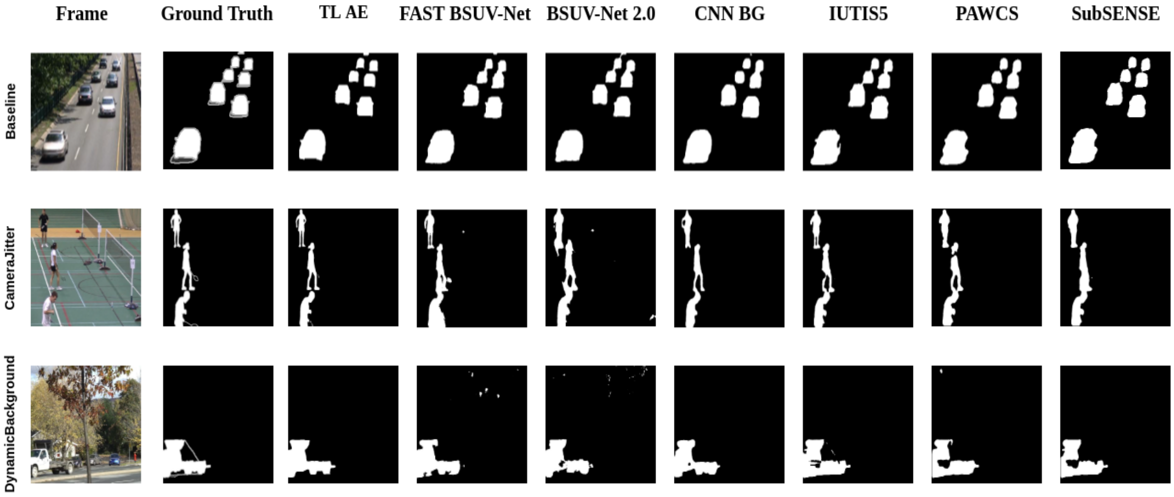


Figure 10. Visual comparison between proposed and reference models generated foreground masks

4. DISCUSSION

The model was trained using a transfer learning technique in the first experiment, while in the second experiment we went dive deeper into auto-encoder architecture for training a multi-depth model. As shown in section 3, the PCA+DA strategy provides significant improvements across all objective metrics over other adopted strategies for the transfer learning and multi-depth approaches. Consequently, we used the models based on these strategies to compare with the popular CNN model [9] and the state-of-the-art methods. From the results obtained in Table 4 and Figure 10, we can see that the results of our model based on (TL AE) achieve improved performance than those obtained by the reference algorithms.

When comparing our two developed approaches, the achieved performance by the transfer learning approach due to the transferred knowledge to the first layers of the encoder part trained with the VGG16 (ImageNet dataset) model. Furthermore, the key information provided by the main components in the PCA transformation is ordered according to their power of representation, which encourages features in the dataset to be statistically independent. The addition of transformed images by PCA at several rates has the objective to increase the training dataset size while preserving essential information. In addition, to provide adapted and meaningful data to recompense for the loss in the first layer of the decoder part of the network. As a result, we prove the effectiveness of the proposed and novel data augmentation strategy. This strategy is based on the preservation of necessary information, which proves its ability to avoid the over-fitting impact.

5. CONCLUSION

Motivated by the recent development of moving object segmentation methods based on deep learning, we presented experiments comparing two deep learning approaches trained using three different strategies to increase data size. The main purpose of the proposed methods is to increase the dataset size using different strategies to improve the model accuracy. The adopted data augmentation strategies are; PCA technique-based geometric transformation and classical data augmentation. The aforementioned strategies were adopted to reduce the over-fitting problem as well as to generate the required features for the moving object segmentation while improving the model performance. The deep learning category used in our experiments is based on a deep convolution auto-encoder, which is mostly used for image segmentation tasks.

The main objective is to enhance the object segmentation method based on a supervised deep auto-encoder using limited training data. However, it can be concluded that combining morphological and geometrical transformation for model training, helps the model enhance its generalization capabilities and generate a precise model with minimal training data. Furthermore, compared to the traditional data augmentation techniques (mirroring, rotation, and shifting) that rely on changing the placement of the coordinates in the same mathematical plane which produces correlated variables and can offer minimal enhancement. Our work demonstrates the value of purposefully enriching training data as with PCA to create a new representation of the variables in a new plane with an important variance. As well as extracting the variables required for the segmentation task and removing the unnecessary variables that can distort the results of the prediction.

REFERENCES

[1] N. Rachburee and W. Punlumjeak, "An assistive model of obstacle detection based on deep learning: YOLOv3 for visually impaired people," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 4, pp. 3434–3442, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3434-3442.

[2] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[3] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, Sep. 2018, doi: 10.1016/j.asoc.2018.05.018.

[4] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016, doi: 10.1016/j.neucom.2015.09.116.

[5] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017, doi: 10.1016/j.neucom.2016.12.038.

[6] Y. Wang, Z. Luo, and P.-M. Jodoin, "Interactive deep learning method for segmenting moving objects," *Pattern Recognition Letters*, vol. 96, pp. 66–75, Sep. 2017, doi: 10.1016/j.patrec.2016.09.014.

[7] J. Gracewell and M. John, "Dynamic background modeling using deep learning autoencoder network," *Multimedia Tools and Applications*, vol. 79, no. 7–8, pp. 4639–4659, Feb. 2020, doi: 10.1007/s11042-019-7411-0.

[8] A. Bouguettaya, H. Zarzour, A. Kechida, and A. M. Taberkit, "Vehicle detection from UAV imagery with deep learning: a review," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2021, doi: 10.1109/TNNLS.2021.3080276.

[9] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognition*, vol. 76, pp. 635–649, Apr. 2018, doi: 10.1016/j.patcog.2017.09.040.

[10] W. Ge, Z. Guo, Y. Dong, and Y. Chen, "Dynamic background estimation and complementary learning for pixel-wise foreground/background segmentation," *Pattern Recognition*, vol. 59, pp. 112–125, Nov. 2016, doi: 10.1016/j.patcog.2016.01.031.

[11] T. Bouwmans, “Traditional and recent approaches in background modeling for foreground detection: An overview,” *Computer Science Review*, vol. 11–12, pp. 31–66, May 2014, doi: 10.1016/j.cosrev.2014.04.001.

[12] H. Liu, X. Han, X. Li, Y. Yao, P. Huang, and Z. Tang, “Deep representation learning for road detection using Siamese network,” *Multimedia Tools and Applications*, vol. 78, no. 17, pp. 24269–24283, Sep. 2019, doi: 10.1007/s11042-018-6986-1.

[13] H. Salehinejad, S. Naqvi, E. Colak, J. Barfett, and S. Valaee, “Cylindrical transform: 3D semantic segmentation of kidneys with limited annotated images,” in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov. 2018, pp. 539–543, doi: 10.1109/GlobalSIP.2018.8646668.

[14] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[15] R. Keshari, M. Vatsa, R. Singh, and A. Noore, “Learning structure and strength of CNN filters for small sample size training,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 9349–9358, doi: 10.1109/CVPR.2018.00974.

[16] H. Salehinejad, S. Valaee, T. Dowdell, and J. Barfett, “Image augmentation using radial transform for training deep neural networks,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 3016–3020, doi: 10.1109/ICASSP.2018.8462241.

[17] D. Bank, N. Koenigstein, and R. Giryes, “Autoencoders,” *arXiv preprint arXiv:2003.05991*, Mar. 2020

[18] S. C. Leonov, A. Vasilyev, A. Makovetskii, V. Kuznetsov, and J. Diaz-Escobar, “An algorithm for selecting face features using deep learning techniques based on autoencoders,” in *Applications of Digital Image Processing XLI*, Sep. 2018, doi: 10.1117/12.2321068.

[19] S. A. Ebiaredoh-Mienye, E. Esenogho, and T. G. Swart, “Artificial neural network technique for improving prediction of credit card default: A stacked sparse autoencoder approach,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 5, pp. 4392–4402, Oct. 2021, doi: 10.11591/ijece.v11i5.pp4392-4402.

[20] K. Seddiki *et al.*, “Towards CNN representations for small mass spectrometry data classification: from transfer learning to cumulative learning,” *bioRxiv*, Mar. 2020.

[21] I. Idrissi, M. Azizi, and O. Moussaoui, “Accelerating the update of a DL-based IDS for IoT using deep transfer learning,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 23, no. 2, pp. 1059–1067, Aug. 2021, doi: 10.11591/ijeecs.v23.i2.pp1059-1067.

[22] M. Al-Smadi, M. Hammad, Q. B. Baker, and S. A. Al-Zboon, “A transfer learning with deep neural network approach for diabetic retinopathy classification,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 4, pp. 3492–3501, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3492-3501.

[23] F. Zhuang *et al.*, “A comprehensive survey on transfer learning,” *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[24] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[25] S. M. Shaharudin, N. Ahmad, and S. M. C. M. Nor, “A modified correlation in principal component analysis for torrential rainfall patterns identification,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 4, pp. 655–661, Dec. 2020, doi: 10.11591/ijai.v9.i4.pp655-661.

[26] C. Kamlaskar and A. Abhyankar, “Multilinear principal component analysis for iris biometric system,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 23, no. 3, pp. 1458–1469, Sep. 2021, doi: 10.11591/ijeecs.v23.i3.pp1458-1469.

[27] C. Clausen and H. Wechsler, “Color image compression using PCA and backpropagation learning,” *Pattern Recognition*, vol. 33, no. 9, pp. 1555–1560, Sep. 2000, doi: 10.1016/S0031-3203(99)00126-0.

[28] P. C. Yuen, “Human face recognition using PCA on wavelet subband,” *Journal of Electronic Imaging*, vol. 9, no. 2, Apr. 2000, doi: 10.1117/1.482742.

[29] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *3rd International Conference on Learning Representations*, Sep. 2014.

[30] A. W. Reza, M. M. Hasan, N. Nowrin, and M. M. Ahmed Shibly, “Pre-trained deep learning models in automatic COVID-19 diagnosis,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 22, no. 3, pp. 1540–1547, Jun. 2021, doi: 10.11591/ijeecs.v22.i3.pp1540-1547.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, 2012, vol. 25.

[32] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, “CDnet 2014: an expanded change detection benchmark dataset,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2014, pp. 393–400, doi: 10.1109/CVPRW.2014.126.

[33] F. Chollet, *Keras: The python deep learning library*. Astrophysics Source Code Library, 2018.

[34] E. Bisong, “Google colaboratory,” in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Berkeley, CA: Apress, 2019, pp. 59–64, doi: 10.1007/978-1-4842-4470-8_7.

[35] T. S. Gunawan *et al.*, “Development of video-based emotion recognition using deep learning with Google Colab,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 5, pp. 2463–2471, Oct. 2020, doi: 10.12928/telkomnika.v18i5.16717.

[36] M. O. Tezcan, P. Ishwar, and J. Konrad, “BSUV-Net: a fully-convolutional neural network for background subtraction of unseen videos,” in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2020, pp. 2763–2772, doi: 10.1109/WACV45572.2020.9093464.

[37] M. O. Tezcan, P. Ishwar, and J. Konrad, “BSUV-Net 2.0: spatio-temporal data augmentations for video-agnostic supervised background subtraction,” *IEEE Access*, vol. 9, pp. 53849–53860, 2021, doi: 10.1109/ACCESS.2021.3071163.

[38] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, “SuBSENSE: a universal change detection method with local adaptive sensitivity,” *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, Jan. 2015, doi: 10.1109/TIP.2014.2378053.

[39] S. Bianco, G. Ciocca, and R. Schettini, “Combination of video change detection algorithms by genetic programming,” *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 6, pp. 914–928, Dec. 2017, doi: 10.1109/TEVC.2017.2694160.

[40] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, “Universal background subtraction using word consensus models,” *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4768–4781, Oct. 2016, doi: 10.1109/TIP.2016.2598691.

BIOGRAPHIES OF AUTHORS

Abdeldjalil Kebir is a Ph.D. Student at the Badji Mokhtar University-Annaba (Algeria) and a member of laboratory Automatic and Signal Processing of Annaba (LASA). He received his BS and MS Degrees (Communication and Digital Processing) from the same institution in 2011 and 2013 respectively. His main research interests include video and image segmentation using recent machine learning methods. He can be contacted at email: kebirabdeldjalil@gmail.com.

Mahmoud Taibi received his BSc from the USTO University-Oran (Algeria) in Electrical Engineering in 1980, then an MSc degree from Badji-Mokhtar University-Annaba (Algeria) in 1996. Currently, he is a full professor in Computer Science since 2006 at Badji-Mokhtar University-Annaba (Algeria). He is a member with the LERICA laboratory. His research interests focus on intelligent systems, intrusion detection, methods used in automatic object detection and tracking systems, as well as techniques and tips used in the field of Deep Learning. He can be contacted at email: mahmoudtaibi@yahoo.fr.