A DEEP-NARMA FILTER FOR UNUSUAL BEHAVIOR DETECTION FROM VISUAL, THERMAL AND WIRELESS SIGNALS

Nikolaos Bakalos[1], Athanasios Voulodimos[2], Anastasios Doulamis[1], and Nikolaos Doulamis[1]

[1] National Technical University of Athens, Athens 15773, Greece
[2] Department of Informatics and Computer Engineering, University of West Attica, Athens 12243, Greece

ABSTRACT

Detection of unusual behavior is an important topic in signal and image processing. Because of the topic's complexity, addressing it as a solely RGB video analysis problem raises significant challenges. This has resulted in approaches that aim at exploiting different data modalities that can overcome the inherent restrictions of unimodal techniques. Moreover, the classification outcome of such approaches is affected not only by the input data, but also by previous classification history. To this end, this paper introduces a novel deep-NARMA filter that extends a typical CNN architecture, and endows it with autoregressive moving average behavior. In addition, it incorporates a data fusion framework that supplements RGB video streams, with thermal capturing and information about the distortion of WiFi signal reflectance. Experimental results indicate a better performance compared to conventional as well as deep learning approaches.

***Index Terms*—**unusual behavior, deep learning, NARMA.

1. INTRODUCTION

Identification of unusual or abnormal behavior is a crucial topic in research community triggering a range of applications in different domains such as security/safety and production quality assurance [1], [2], [3]. Nowadays, the most popular way to identify unusual human behaviors is through the exploitation of video cameras (either disjoint or overlapping) on the use of video surveillance systems. Towards this direction, computer vision tools and machine learning algorithms are properly interwoven to detect physical intrusions, especially of humans, usually for public area or critical infrastructures [4]. The main drawback of all these single modality approaches, are the inherent restrictions of the information captured. For instance, approaches based on unimodal visible spectrum analysis are vulnerable to occlusions, luminosity changes, light reflections, etc. This has recently spawned a number of techniques that leverage multimodal processing and data fusion, to compensate these inherent limitations by supplemental information from other modalities.

Apart from the Red Green Blue (RGB) visible spectrum data, thermal information is another useful input for detecting human intrusion or unusual behaviors. Thermal sensors are not sensitive to changes in illumination [5]. However, the information captured does not include texture or color information, and due to the fact that target objects are not always homogeneous in temperature, object detection becomes an arduous task. Since both RGB and thermal sensing are actually based on visual cues, they should be supplemented by additional data that are not limited by the restrictions of visual information (e.g. occlusions). An interesting modality for consideration, is the monitoring and analysis of distortions in radiofrequency transmissions such as reflectance of WiFi signals, used for wireless communication [6]. However, unusual behavior identification cannot be accurately produced by simply monitoring WiFi reflectance, as they cannot model complex behavior (e.g., human motion trajectories), thus the modality is mainly useful as additional information combined with other information streams, (i.e., thermal and RGB imaging). For this reason, information fusion across the aforementioned modalities improves unusual behavior detection.

The main difficulty in such fusion process is to extract a suitable set of features that can represent the meanings of the fused complex data. Convolutional Neural Networks (CNNs) have shown to be excellent feature detectors [7]. For this reason, our classifier is based on a CNN framework. Nevertheless, the key problem of adopting conventional CNN architecture in our case is that spikes may appear in the classification outputs since input data are independently processed per time instance. To address this, we need to modify the traditional CNNs so as to be able to model auto-regressive and moving average (ARMA) ([9],[10]) behaviors. This allows for a smooth human intrusion detection process since previous classification outputs are also taken into account for the classification of current states. Thus, we need to introduce a modification of the CNN architecture to capture the ARMA properties in a non-linear context.

1.1 Related Work

Detection of unusual behavior is a well-studied problem of smart surveillance such as in [11] and [12]. A definition of abnormal behavior is given in [1] as an occurrence of abnormal events that are *"rare in the scene and which are*

different from the majority". Smart surveillance systems process video streams using a variety of techniques such as pixel based techniques [13], trajectory based techniques [14], a fusion of trajectory and pixel information [1], object detection frameworks such as [15], background and target modelling [16], object tracking [17], activity recognition [18], [19], crowd dynamics [20], real-time critical application scenarios [21] and even choreographic time series modelling [22]. These techniques though powerful, leverage information relevant to visual cues, thus they are vulnerable to occlusions, hidden fields of view or poor visual conditions. Thermal information is also considered as an alternative [5], but with the limitation of low representation capabilities in texture modelling.

On the other hand, localization by using properties of radiofrequency devices can also provide useful information. These techniques are either device free approaches based on Software Defined Radio (SDR) or approaches that leverage commercial off the shelf (COTS) devices [6]. The device free approach analyses the Received Signal Strength of a transmitted signal but they do not provide efficient accuracy [23]. On the contrary COTS approaches have been shown to provide accurate detection of human presence, since they exploit Channel State Information (CSI) [9].

Recently, deep learning techniques, such as CNN, have been favored over traditional learning methods, such as SVM or shallow feedforward neural networks, for unusual behavior detection [1], [18], [7]. This is mainly due to high representational capabilities of deep learning methods. However, the limitation of such approaches is that they do not take into consideration previous classification outputs that need to be fed back to the network to model an autoregressive behavior.

1.2 Our Contribution

In this paper, we introduce a Deep-Non-linear Auto-Regressive Moving Average (Deep-NARMA) filter that leverages the representational capabilities of Convolutional Neural Networks (CNN), properly modified to cope with the autoregressive nature of a tapped delay line, and transforms the inputs in an efficient non-linear feature map. The proposed classifier achieves an effective feature representation of the heterogeneous inputs but it simultaneously introduces an input-and output memory. We also propose a novel data fusion processing of three different modalities to improve the detection accuracy of unusual behavior. Specifically, alongside the normal RGB surveillance, we leverage also thermal imaging, as well as measurements of wireless signal reflection for human presence detection. While a number of works has been published using fusions of RGB and thermal modalities, such as in [24] and [25], to our knowledge there are no previous works in the literature that consider fusion of thermal and RGB imaging with WiFi reflectance data.

2. THE PROPOSED DEEP NARMA FILTER

Let us denote as $y(n) = [P_s, P_u]^T$ a *2x1* vector that contains the probabilities P_s and P_u, that the observations at time instance n can be classified as suspicious or unusual behavior (class s) or normal behavior (class u). Let us now assume that there is a non-linear function that relates probabilities $y(n)$ with some measurable observations $x(n)$. The output $y(n)$ is related with the current and previous observations over a time window of q previous instances. We denote as $x(n-j), j = 0, \dots, q-1$ these q previous observations. In the following notation, we assume that $x(n)$ are multidimensional tensors of the input data. Assuming a non-linear dependency of the classification output and the previous classification values, we derive a non-linear autoregressive-moving average model:

$$y(n) = g\big(x(n-1), \dots, x(n-q), y(n-1), \dots, y(n-p)\big) + e(n) \quad (1)$$

where $g(\cdot)$ refers to the non-linear relationship and p, q express the order of the model. Vector $e(n)$ is an independent and identically distributed error. Eq. (1) cannot be easily calculated, as $g(\cdot)$ is unknown. It is clear that Eq. (1) resembles an NARMA(p,q) model. The use of machine learning methods can produce an approximation of $g(\cdot)$ in a way that minimizes the error $e(n)$. A feed forward neural network (FNN) with a tapped delay line input filter can simulate the behavior of a NARMA model [26]. A recursive implementation has also been proposed in [27]. However, this FNN model fails at effectively selecting features of high-dimensional space and complex heterogeneous environments.

Convolutional Neural Networks have demonstrated excellent representational capabilities in feature selection such as in [7], [8], [28]. The proposed Deep-NARMA filter combines the effectiveness in feature selection of CNN with the autoregressive nature of a tapped delay line, in order to select optimal features that enable the classification of the observed behaviors. We extend the traditional CNN architecture by adding a tapped delay line in the input layer, to organize the external input data $x(n)$ as well as to feed back the previous classification outputs. This extension includes two terms, the moving average term that delays the input $x(n)$ for q discrete previous times, and the autoregressive term that delays the output $y(n)$ for p previous discrete instances. Previous classification outputs affect the current output, as temporal dependencies occur.

After the expanded input layer that receives the current data (RGB, thermal, WiFi) and the delayed responses over previous times, we proceed with the convolutional/pooling layers. This layer applies convolutional transformations on the input data so as to maximize the classification performance. The convolutions are executed over the input data and a set of kernels, in order to select appropriate features. The kernel parameters are estimated in a way that minimizes the performance error on a ground-truth training

set. The L feature maps, denoted as $f_1, f_2, \ldots, f_L$ are used as inputs in the final (classification) layer. In the experimental evaluation of section 4, the convolutional/pooling layers consist of three different convolutional layers, with 5x5x4, 5x5x32 and 5x5x32 respective filter sizes, separated by the ReLU and Max pooling components.

The final component of the filter is the classification layer that receives the $f_1, f_2, \ldots, f_L$ feature maps and triggers a supervised behavior classification. The f_i feature maps are tensors with dimensions that express the spatial attributes and the different modalities of the input data.

The classification layer consists of r neurons, each stimulating a non-linear operation, where the sigmoid is neuron activation function. If we denote as $w_{i,j}$ the weights that connect the i-th feature map f_i with the j-th hidden neuron of the classification layer, then the output of this neuron will be $u_j = \varphi(\mathrm{w}_j^T \cdot f)$, where f is the aggregate feature map concatenating all features f_i and w_j the aggregate weights for the j-th hidden neuron. Then, output will be given as:

$$y_w(n) = \varphi\left(v^T \cdot u\right) \equiv \varphi(z_w(n)) \qquad (2)$$

where u includes all outputs u_j over all the r hidden neurons and v the aggregate weights connecting the r hidden neurons of the classification layer with the output neuron. In Eq. (2), $z_w(n)$ expresses the input of the final output neuron before applying the activation function $\varphi(\cdot)$. In the previous notation, we have assumed that the classification output consists of one neuron. Extension to multiple neurons is straightforward. Subscript w in Eq. (6) denotes the dependence of the classification on the network weights which will be estimated through a learning process. In our configuration, the proposed model consists of 64 hidden layers and two output neurons.

A schematic of the proposed architecture is presented in figure.

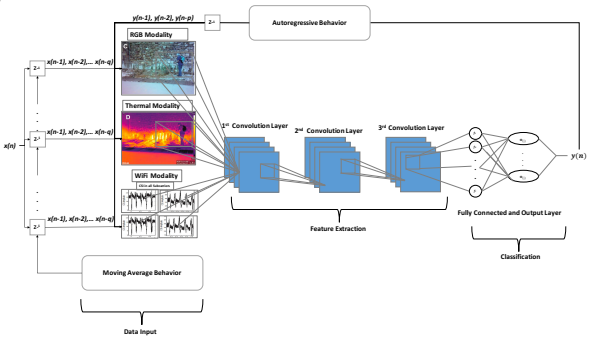

Figure 1: Architecture of the proposed Deep-NARMA Filter

3. MODELLING OF INPUT DATA

3.1 Visual Spectrum Imaging (RGB)

The most common surveillance data modality consists of RGB video streams. The raw captured data are preprocessed using the YOLO (You Only Look Once) object detection framework [15], which models the object detection as a regression problem by separating the input image in bounding boxes, that are assigned class probabilities. The object detection is executed via a CNN architecture of 24 convolutional layers and 2 fully connected layers.

Each frame is described as a class image CL_{RGB}, with the same size as the RGB image, where the *(x,y)* pixel of the RGB image *I(x,y)* is denoted as $o_{k,RGB}(x,y)$, in the class in the following way:

$$CL_{RGB}(x,y) = o_{k,RGB}(x,y) \qquad (3)$$

where k represents the object with identity k in the YOLO object detection framework.

3.2 Thermal Imaging

The thermal captured data are preprocessed using the background subtraction algorithm of [5]. Similarly with the RGB modality, we extract a class label image CL_T, the same size as the input thermal frame T, where the *(x,y)* pixel of T is denoted in the class label image as:

$$CL_T(x,y) = o_{b,T}(x,y), b = \{Background, Foreground\} \quad (4)$$

3.3 WiFi signal reflection

The identification of human intrusion in this modality is achieved by exploiting the Channel State Information (CSI) metric of commercial WiFi devices, as is the case in [9]. CSI models the signal propagation from a transmitter to a receiver, and supports many subcarriers, due to the Orthogonal frequency Division Multiplexing (OFDM) principle. Physical attributes of the wireless channel (e.g. power decay, scattering) are measured with respect to distance, fading, shadowing and effects of interference [9] The measurement of the K available sub-carriers is:

$$H(n) = [H(n,f_1)\ H(n,f_2)\ \cdots\ H(n,f_K)]^T \qquad (5)$$

where $H(n,f_i)$ refers to the amplitude and the phase of the i-th subcarrier with central frequency f_i. Therefore, we have that: $H(n,f_i) = |H(n,f_i)|e^{j\angle H(n,f_i)}$.

A preprocessing of $H(n)$ must be executed to remove outliers and noise. This is achieved via the use of a Hampel identifier and wavelet denoising respectively [10]. After removing outliers and noise we proceeded in normalizing the signals and eigenvector processing. The final stage of preprocessing includes normalization, correlation of subcarriers and eigenvector processing of the signals (Fig. 2). The pre-processed CSI data are analyzed using a linear SVM classifier in order to detect human intrusions in a scene. The produced classification IDs, say $C_{WiFi}(n)$ are also used as input in the proposed fused deep learning classifier.,

$$x_{wifi}(n) = [H(n)\ C_{WiFi}(n)]^T \qquad (6)$$

4. EXPERIMENTAL VALIDATION

4.1 Dataset Description

The evaluation of the proposed solution used a dataset that has been captured as part of the EU Horizon 2020 STOP-IT

Project (https://stop-it-project.eu/) (grant agreement No. 740610), a research initiative that addresses the protection of critical water infrastructure. The dataset consists of RGB, and thermal video streams and WiFi reflectance information. The RGB data were captured using and OB-500Ae camera with 1280x720 pixel resolution and 30 fps framerate, while for the thermal capturing we used a Workswell InfraRed Camera 640 (WIC) with a 640x512 pixel resolution and a 30 fps framerate. The WiFi reflectance data were captured using a transmitter-receiver couple that consisted of a WiFi router (TP-Link N300 TL-WR841N) and an Intel 5300 NIC receiver, capturing once every 10 seconds. The data were labeled based on pre-determined scenarios co-defined by end users that designated whether the captured human behavior over all data modalities can be considered as unusual. All the data were normalized to be in the same range, from 0 to 1.

The computer used for all training and testing was an Intel® Core™ i7-6700 CPU@ 4000 GHz CPU with 16GB of RAM and an NVIDIA GeForce GTX 1070 with 8GB DDR5 memory. The deep learning models also used the CUDA 9.2 Toolkit.

4.2 Evaluation Results

The classification performance of the proposed Deep Convolutional NARMA filter was compared with different classifiers. To illustrate the results of fusion between different modalities we executed unimodal classification scenarios with (i) a linear kernel SVM, (ii) a non-linear Radial Basis Function SVM (RBF-SVM), (iii) 2 different architectures of a traditional feedforward neural network with 1 hidden layer of 10 neurons/layer and 2 hidden layers of 10 neurons/layer respectively. The results of these models in the unimodal and the multimodal case, are presented in Figure 2, which showcases how data fusion of the proposed modalities increased the classification accuracy.

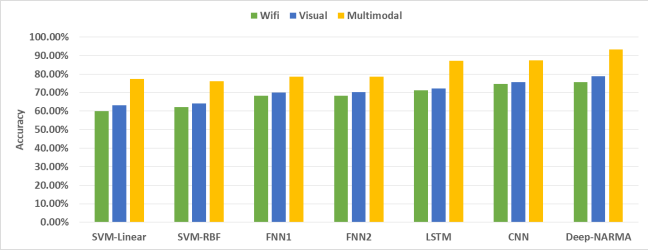

Figure 2: Effect of fusion of different data modalities in classification performance

In the multimodal use case, we proceeded in extensive evaluation of the proposed model, comparing it with the aforementioned "shallow" classifiers, as well as with two additional deep learning models, a Long Short-Term Memory (LSTM) deep recurrent neural network, as well as with a CNN. The CNN's structure was identical with the one used in the proposed Deep NARMA filter, but without the proposed expansion to generate the model's autoregressive behavior. The classification performance was measured using traditional performance metrics, i.e. accuracy, precision, recall and F1 scores. The results of this evaluation is depicted in Table 1. Moreover, the effects of adding autoregressive characteristics to the model are presented in Figure 3

Table 1: Classification performance metrics on multimodal experiments

Classification Method	Precision	Recall	Accuracy	F1 Score
SVM-Linear	68.51%	61.71%	77.36%	64.93%
SVM-RBF	66.99%	60.06%	76.11%	63.34%
FNN1	69.95%	63.30%	78.52%	66.46%
FNN2	70.13%	63.50%	78.66%	66.65%
LSTM	81.14%	76.12%	87.11%	78.55%
CNN	81.62%	76.69%	87.46%	79.08%
Deep-NARMA	89.76%	86.66%	93.23%	88.18%

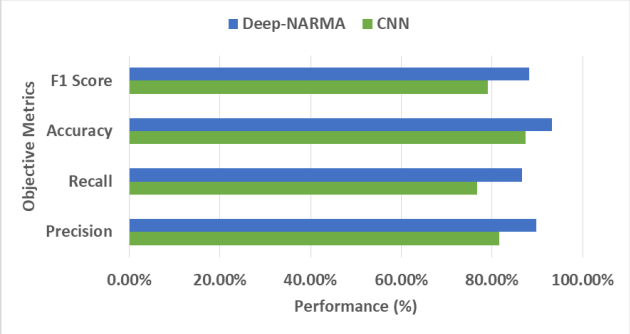

Figure 3: Effect of autoregressive behavior in the classification performance

5. CONCLUSION

The identification of unusual behavior from surveillance systems is an arduous and complex task, the performance of which is bounded by the model's properties and the inherent limitations of the captured data modality. To address these drawbacks we proposed a deep convolutional NARMA model alongside a multimodal data fusion framework, that expands the monitoring capabilities of the model beyond a single data type, and allows it to better adapt to dynamic events, such as suspicious human movements.

The proposed methods were experimentally evaluated using a dataset captured in the context of the Horizon 2020 STOP-IT project. The results clearly indicate that autoregressive and multimodal approaches enhance typical deep learning models in terms of performance for identifying unusual behaviors.

6. ACKOWLEDGEMENT

The research leading to these results has received funding from the EU H2020 research and innovation programme under grant agreement No. 740610, project STOP-IT.

7. REFERENCES

[1] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L.O Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance." *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[2] D.I., Kosmopoulos, N.D., Doulamis, A.S., Voulodimos, "Bayesian filter based behavior recognition in workflows allowing for user feedback," *Computer Vision and Image Understanding*, 116 (3), pp. 422-434, 2002.

[3] K. Makantasis, A. Doulamis, N. Doulamis and K. Psychas, "Deep Learning Based Human Behavior Recognition In Industrial Workflows," IEEE Internaltional Conference on Image Processing, (ICIP), Arizona, USA, Sept. 2016.

[4] V. Sze, Y.H. Chen, J. Emer, A. Suleiman, and Z. Zhang "Hardware for machine learning: Challenges and opportunities." IEEE Custom Integrated Circuits Conference (CICC) pp. 1-8, 2017.

[5] K. Makantasis, A. Nikitakis, A. Doulamis, N. Doulamis and Y. Papaefstathiou, "Data-Driven Background Subtraction Algorithm for in-Camera Acceleration in Thermal Imagery," in IEEE Transactions on Circuits and Systems for Video Technology, 2017.

[6] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," ACM SIGCOMM Comput. Commun. Rev., vol. 41, no. 1, p. 53, 2011.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] N. Doulamis and A. Doulamis, "Semi-Supervised Deep Learning for Object Tracking and Classification," *IEEE International Conference on Image Processing (ICIP)*, pp. 848-852, Paris, France, 2014.

[9] Hai Zhu, Fu Xiao, Lijuan Sun, Ruchuan Wang, and Panlong Yang, "R-TTWD: Robust Device-Free Through-The-Wall Detection of Moving Human with WiFi", IEEE Journal on selected areas in communications, vol. 35, no. 5, May 2017.

[10] L. Davies and U. Gather, "The identification of multiple outliers," *J. Amer. Statist. Assoc*., vol. 88, no. 423, pp. 782–792, 1993.

[11] O. Popoola and K. Wang, "Video-based abnormal human behavior recognition -a review," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,vol. 42, no. 6, pp. 865–878, Nov 2012.

[12] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 3, pp.367–386, March 2015

[13] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2010, pp. 1975–1981

[14] K. Ouivirach, S. Gharti, and M. N. Dailey, "Incremental behavior modeling and suspicious activity detection," Pattern Recognition, vol. 46, no. 3, pp.671–680, 2013. [Online]. Available:http://www.sciencedirect.com/science/article/pii/S0031320312004426

[15] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE CVPR, Las Vegas, NV, 2016, pp. 779-788.

[16] S. Herrero and J. Bescs. "Background subtraction techniques: Systematic evaluation and comparative analysis," *11th International Conference on Advanced Concepts for Intelligent Vision Systems, ser. ACIVS '09*. Springer-Verlag, 2009.

[17] D. S. Yeo, "Superpixel-based tracking-by-segmentation using markov chains.," *IEEE Conference in Computer Vision and Pattern Recognition (CVPR),* 2017.

[18] D. Kosmopoulos, A. Voulodimos, A. Doulamis, "A system for multicamera task recognition and summarization for structured environments," *IEEE Transactions on Industrial Informatics*, 9 (1), 161-171, 2013.

[19] A.S Voulodimos, D.I Kosmopoulos, N.D, Doulamis, T.A Varvarigou, "A top-down event-driven approach for concurrent activity recognition," *Multimedia Tools and Applications,* 69 (2), pp. 293-311, 2014.

[20] H.M. Mousavi, "Analyzing tracklets for the detection of abnormal crowd behavior," *IEEE Winter Conference on In Applications of Computer Vision (WACV),* 2015.

[21] A. Doulamis, N. Doulamis, E. Protopapadakis and A. Voulodimos, "Combined Convolutional Neural Networks and Fuzzy Spectral Clustering for Real Time Crack Detection in Tunnels," *25th IEEE International Conference on Image Processing (ICIP)*, Athens, 2018, pp. 4153-4157.

[22] I. Rallis, N. Doulamis, A. Voulodimos and A. Doulamis, "Hierarchical Sparse Modeling for Representative Selection in Choreographic Time Series," *25th IEEE International Conference on Image Processing (ICIP)*, Athens, 2018, pp. 1023-1027, 2018.

[23] K. Wu, J. Xiao, Y. Yi, M. Gao, and L. M. Ni, "FILA: Fine-grained indoor localization," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2210–2218.

[24] D. Jiang, D. Zhuang, Y. Huang and J. Fu, "Survey of multispectral image fusion techniques in remote sensing applications", *Image Fusion and its applications*, Y. Zheng, INTECH Open Access Publisher, Vol. 1, pp. 1-22, 2011.

[25] A. R. Pal and A. Singha, "A comparative analysis of visual and thermal face image fusion based on different wavelet family," *2017 International Conference on Innovations in Electronics, Signal Processing and Communication (IESC),* Shillong, 2017, pp. 213-218.

[26] J. Connor, D. Martin, and L. Altas, "Recurrent neural networks and robust time series prediction," IEEE Trans. Neural Networks, Vol. 5, pp. 240–254, Mar. 1994.

[27] A.D. Doulamis, N.D. Doulamis, and S. D. Kollias, "An adaptable neural-network model for recursive nonlinear traffic prediction and modeling of MPEG video sources," IEEE Transactions on Neural Networks, Vol. 14, No. 1, pp. 150-166, 2003.

[28] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," Computational Intelligence and Neuroscience, vol. 2018, Article ID 7068349, 13 pages, 2018.