Convolutional neural networks for the analysis of broadcasted tennis games

Grigorios Tsagkatakis [†]***, Mustafa Jaber*** [⋆]***, Panagiotis Tsakalides*** [†‡]

[†] ***Institute of Computer Science, Foundation for Research & Technology - Hellas (FORTH), Crete, 73100, Greece.***

[⋆] ***NantOmics, LLC, Culver City, CA, 90230, USA.***

[‡] ***Department of Computer Science, University of Crete, Crete, 73100, Greece.***

Abstract

The analysis of complex structured data like video has been a long-standing challenge for computer vision algorithms. Innovative deep learning architectures like Convolutional Neural Networks (CNNs), however are demonstrating remarkable performance in challenging image and video understanding tasks. In this work we propose a architecture for the automated detection of scored points during tennis matches. We explore two approaches based on CNNs for the analysis of video streams of broadcasted tennis games. We first explore the two-stream approach, which involves extracting features related to either pixel intensity values via the analysis of grayscale frames or the encoding of motion related information via optical flow. However, we explore the case of using higher order 3D CNN for simultaneously encoding both spatial and temporal correlations. Furthermore, we explore the late fusion of the individual stream in order to extract and encode both structural and motion spatio-temporal dynamics. We validate the merits of the proposed scheme using a novel manually annotated dataset created from publically available videos.

Introduction

Analysis and understanding of video sequences of complex actions and events has been a long standing challenge for computer vision algorithms. This challenge can be mainly attributed to two aspects of this problem, namely the visual diversity of contextually similar actions and events and the scale and complexity associated with massive video datasets. In this work, we consider the problem of automated detection of winning shots in tennis matches from broadcasted videos.

The primary challenge in this setting is the recognition of complex contextual events from raw pixel values. More specifically, a winning shot in tennis is a shot which leads to (i) a return shot on the net, (ii) a return shot where the ball touches the ground outsize the playing region and (iii) the altogether non-contact of the ball with the opponents' racket, a "winner". Furthermore, broadcasted videos of tennis games are also characterized by varying viewing conditions including overview shots from one side of the court, panoramic shots, zooming to a specific player, replays and others directive choices. Due to these reasons, the automated understanding the contextual information related to a winning shot is a very challenge task.

Despite these challenges, innovative architectures like deep learning, are displaying remarkable performance [1]. Convolutional Neural Networks (CNNs) [2], a particular deep learning architecture, have shown great promise in static image analysis task like object recognition and classification [3, 4] while more recently, they have been introduced in remote sensing [5, 6, 7] and other applications.

Compared to static image understanding, significantly less effort have been given to the analysis of video sequences. Analyzing unstructured video streams is a challenging task for a number of reasons. First, real world dynamics that are manifested in the corresponding video streams, such as changes in viewpoint, illumination and zoom. In addition, while many annotated image datasets are publicly available, a smaller number of labeled datasets is available for videos. Furthermore, analyzing massive, high dimensional video streams is extremely computational demanding, requiring significant resources [8]. Extensions of CNN representations to action recognition in video have been proposed in several recent works [9, 10, 12, 8, 13, 14, 6].

In this work, we consider the problem of detecting winning shots in broadcasted tennis game videos. They key challenges lies in the ability to understand when specific spatio-temporal pattern are associated with contextual events, like a wining ball. We formulate the problem as a binary classification of short video sequences which are encoded though a spatio-temporal deep learning features. The key novelties of this work include:

- Develop a novel dataset for event detection in sports video and more specifically, for winning shots detection in tennis games;
- Investigate 3D Convolution Neural Networks for extracting spatio-temporal features in video sequences;
- Explore the use and fusion of both pixel intensities and motion related optical flow are input data;
- Produce a novel manually annotated dataset for winning shot detection in broadcasted tennis games;
- Demonstrate that accurate detection can be achieved from a limited number of labeled examples.

Related Work

The case of video understanding has been gaining attention while large scale datasets are becoming publically available [8]. We can identify two major approaches, single-stream frame-based methods and two-stream motion-based methods. In the former case either extract spatial features from individual frames which are concatenated for capturing temporal dependencies [15], or they encode spatio-temporal infomration [13, 11]. In the two-stream approaches, both single frame spatial information as well as inter-frame motion descriptions like optical flow are jointly modeled [19, 12]. Encoding temporal information can also be achieved through the use of Long-Short Term Memory (LSTM)

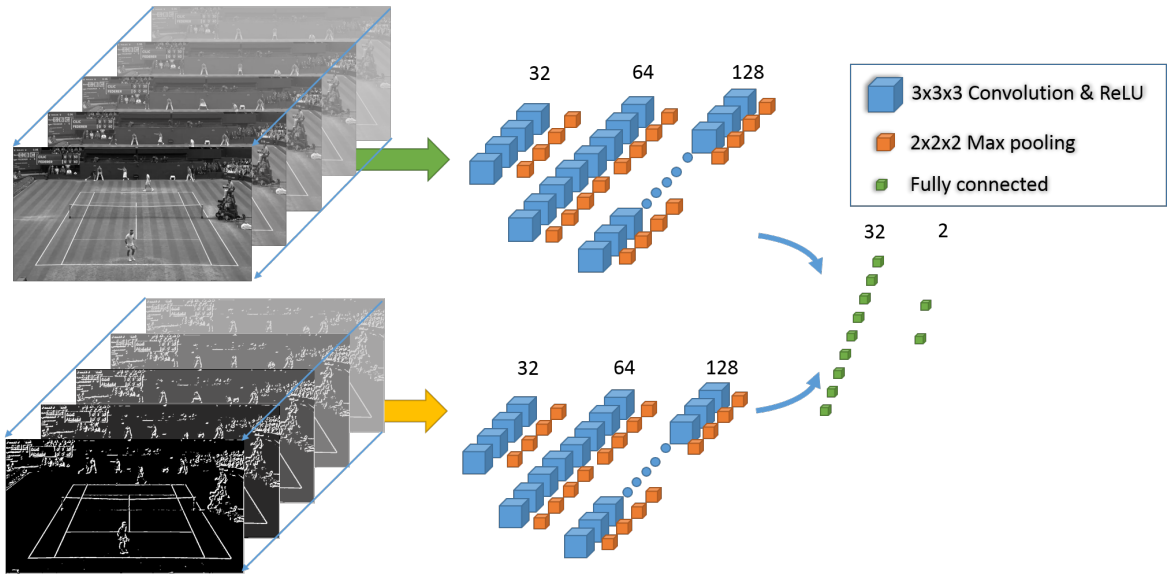

Figure 1. *Block diagram of the proposed architecture. A sequence of frames is introduced as pixel values (top) and optical flow (bottom) to a multi-layer Convolutional Neural Network. The CNN is composed of 2 or 3 concatenated groups of 3D convolution, followed by ReLU activation and max pooling layers while the last layers are fully connected layers. The extracted feature from each 3D CNN are fused together in order to extract structural and motion related features which are mapped to the output layer.*

networks [16], while another concept involves the generation of dynamic images through the collapse of multiple video frames and the use of $2D$ deep feature exaction on such representations [17]. Unlike approaches which consider a single or multiple-frames, two-stream networks consider two source of information, raw frames and optical flow, which are independently encoded by a CNN and jointly classified using methods like Support Vector Machines [6]. Another line of research explore the use of 3D spatio-temporal cubes for encoding actions. An extensive evaluation of 3D CNNs for video classification was recently presented [11].

Proposed model

In this work we explore state-of-the-art CNN based architectures for the analysis of video streams of broadcasted tennis games. The initial point of this work is largely based on the use of 3D CNN for the analysis and classification of video sequences [11]. We extend this mode base considering the three-dimensional extension of the two-stream approach which involves simultaneously extracting features related to both grayscale pixel values and optical flow based data. A key novelty of this work is that the proposed model can encode the dynamics of both structural information in the pixel value domain and motion in the optical flow based domain.

The first type of input data we consider is pixel intensity values were each color frame is first converted to grayscale such that sequences of frames correspond to a given example form a 3D spatio-temporal cube which is introduced to the proposed architecture. Due to computational limitations, the images are resized to 140×78 pixels and we consider temporal sequences of 30 frames corresponding to 1 seconds of video. Figure 1 presents an illustration of the key components of the proposed architecture applied to pixel intensity and optical flow values. To introduce the temporal aspects, we employ the Lukas-Kanade optical flow estimation

In terms of the learning architecture, a CNN is a form of Deep Neural Network, which comprises of convolutional layers alternating with activation and subsampling (pooling) layers, resulting in a hierarchy of increasingly abstract features. At the final layers, fully connected layers are introduced for mapping the features to the specific classes.

Convolutional layer: In typical image recognition tasks 2D convolutions are employed in order to capture shift-invariance in the spatial domain. Extension of the operation to 3D where two encode spatial information and one temporal temporal in similarly introduced for video analysis in order to simultaneously capture spatial and temporal domain invariances Formally, given a $\mathbf{X}$ spatio-temporal cube of dimensions and a filter kernel $\mathbf{w}$ of size $(m \times m \times m)$, the output of the convolutional layer $\mathbf{h}^k$ at spatial location (i, j) and time instance (u) is given by:

$$\mathbf{h}^k_{iju} = (\mathbf{W}^k * \mathbf{x})_{iju} + b^k \tag{1}$$

where b^k is the additive bias term. The key parameters of the convolutional layer is related to the size and number of filters that are learned at each layer. In this work, we follow the approach in [11] and select $32, 64$ and 128 filter of size $3 \times 3 \times 3$.

Non-linear activation: A significant component of a CNN is the non-linear activation associated with the outputs of the convolution. Historically, the two most prominent non-linear function have been the sigmoid and the tanh. However, in recent year, a much simpler function have been introduced. The Rectified Linar Unit (ReLU) preserves the non-negative components while the negative values are set of zero according to: $\hat{\mathbf{x}} = max\{0, \mathbf{x}\}$.

Pooling: The activations of the convolutions are intoduced

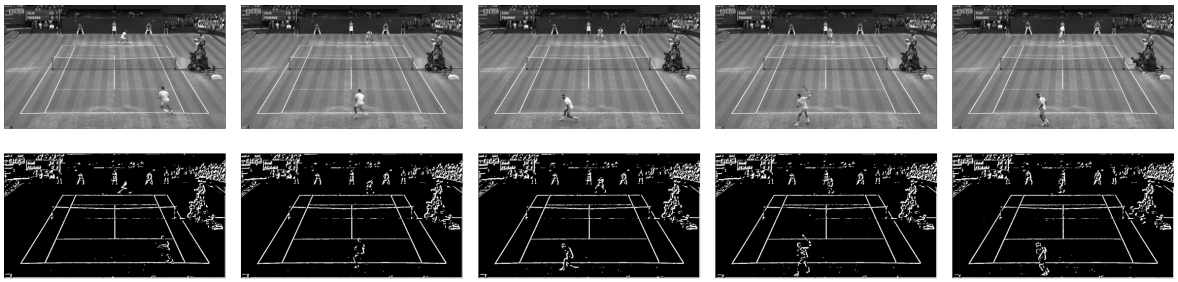

Figure 2. *Exemplary frames of a* ***winning*** *shot in grayscale pixel intensity values (top row) and the associated optical flow values (bottom row).*

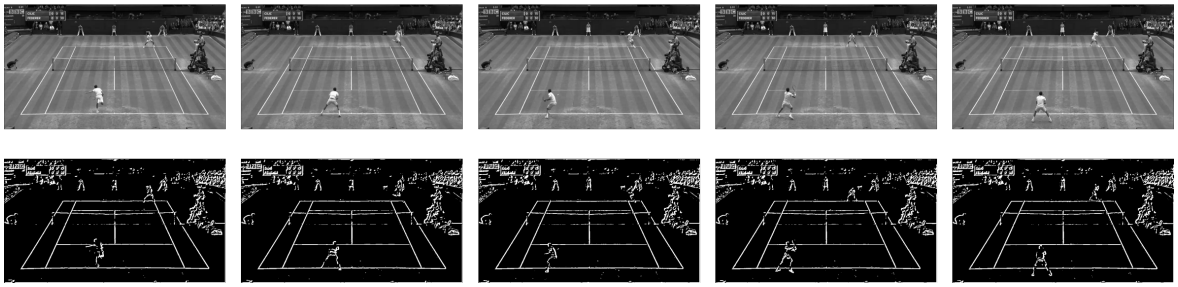

Figure 3. *Exemplary frames of a* ***no-winning*** *shot in grayscale pixel intensity values (top row) and the associated optical flow values (bottom row).*

to a pooling layer that produces downsampled versions of the input maps. In our work, we consider max pooling althrough other methors like average pooling have also been explored.

Fully connected The objective of the fully connected layers layers is to map the features to the specific class. The number of fully connected layers and the candinality of each are crousial parameter both due to the impact the performance and the dramatic increase in computational resources needed. While for datasets like the UCF101 [18], the network much predict hundreds of classes, in our case only two are need. As such, we selected on fully connected layer with 16 neuron. The optimization of the weights and filters took place using the Stochastic Gradient Descent approach with learning rate 10^{-3} with decay 10^{-2}.

Pretrained features

In addition to the features learned during training, we also explored the case of using pre-trained pixel based features extracted from each frame and then introducing 1D convolution for capturing the temporal dependencies. For the $2D$ feature extraction networks, we consider the VGG-16 CNN architecture, which is composed of 13 convolutional layers, with five of them being followed by a max-pooling layer, leading to three fully connected layers [3]. We employed a pretrained model where the weights were estimated through training on thousands of ImageNet database images. For each frame of each sequence, a 4096 dimensional vector is extracted ('fc7')

Dataset

To evaluate the performance of the proposed method, a new dataset was created containing short duration video sequence of winning and non-winning shots. To produce the training and testing example, we considered the final of the 2017 Wimbledon grand slam tournament between Roger Federer and Marin CiliC. The game was broadcasted by BBC and the video is available online on YouTube [1]. Exemplary frames and associated optical flow images from "winner" and "no-winner" sequences are shown in Figures 2 and 3 respectively. One can easily notice that a significant benefit of broadcasted tennis games in contrast to other broadcasted sports is the relatively fixed and static viewpoint, typically from a camera mounted behind on the opponents. For each class, "winners" and "no-winners" sets of 100 example sequence were acquired and we report the performance in terms of classification error. The code utilizes the Keras library with Tensorflow backend for implementing the architectures while the training process was carried on an NVidia K2200 GPU equipped with 640 cores and 4 GBytes of memory [2].

Experimental results

We compile a dataset of high quality broadcasted tennis game. We measure the accuracy of the detection by (i) measuring the detection true positive and false positive rate and (ii) calculating the temporal distance between the time of wining point detection and the instance were the on-screen score is updated. In this work, we report the classification perforation reached after 100 epochs due to both limitation in computational resources but also because no significant gains in performance are observed given the limited training data. To introduce the motion related features, we employ the Lukas-Kanade optical flow estimation with threshold for noise reduction set to 0.009. To make the optical flow compatible with the pixel values, a linear normalization to the $[1, 256]$ range is performed.

We first explore the impact of having deeper CNN by considering two architectures, namely architecture with $[32, 64]$ and $[32, 64, 128]$ convolutional filters followed by 16 and 32 fully con-

[1] https://youtu.be/N4YJ06z5nuk

[2] Code will become available at: https://github.com/spl-icsforth/

nected neuron in the last layer. Figure 4 and Figure 5 present the accuracy achieved using pixel intensity and optical flow as inputs to the CNNs

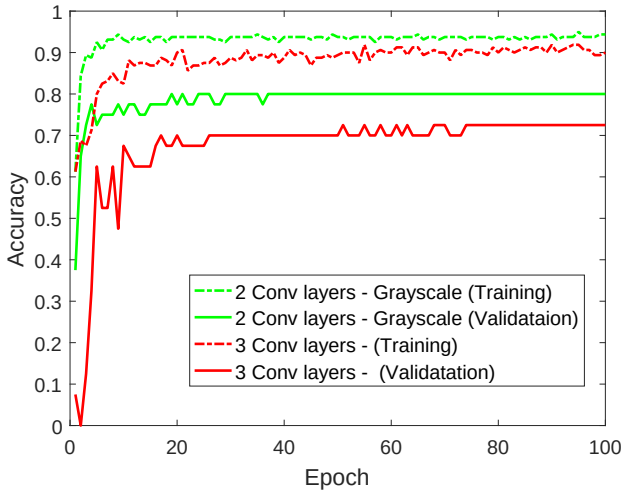

Figure 4. *Accuracy achieved by a 2 and a 3 Convoluational plus FC layer architecture on pixel intensity values.*

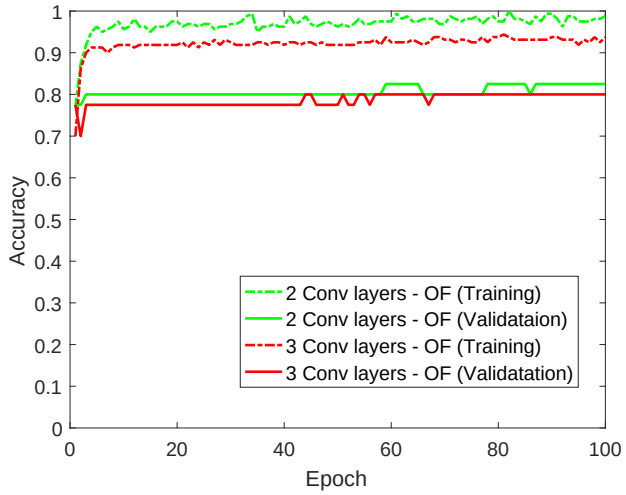

Figure 5. *Accuracy achieved by a 2 and a 3 Convoluational plus FC layer architecture on optical flow values.*

Experimental results presented in Figures 4 and 5 provide key insights into the behavior of the CNN architecture for the task in question. First, we observe that for all cases, the accuracy increases with more epoch since the network is becomes better adapted to the data. However, one can also observe that limited performance gains are observed when reaching the 100 epochs milestone.

Comparing the achieved accuracy at 100 epochs with respect to network depth, we observe that for both pixel intensity and optical flow, the two layer architecture achieves better performance compared to the three layer case, especially for the case of pixel intensities. This may seem counter-intuitive since deeper networks typically perform better. However, in our case, the limited number of training examples and number of classes. In practice this means that the number of network parameters like filters and weights is bigger compared to the amount of training data used. As a result. shallower architectures are more capable in capturing sptio-temporal dynamics. Comparing the performance with respect to the type of input data, we observe that motion related information encoded in optical flow offer significantly better accuracy compared to pixel intensities. Furthermore, it is able to attain this performance from a very small number of training epochs compared to the pixel intensity case.

The reults in Figures 4 and 5 assume that all 200 example sequences are available while during training a $8:2$ split of training/validation with random shuffling per epoch is employed for measuring performance. In Figure 6, we present the impact of the number of training examples for the two and three convolutional and pooling layer architectures.

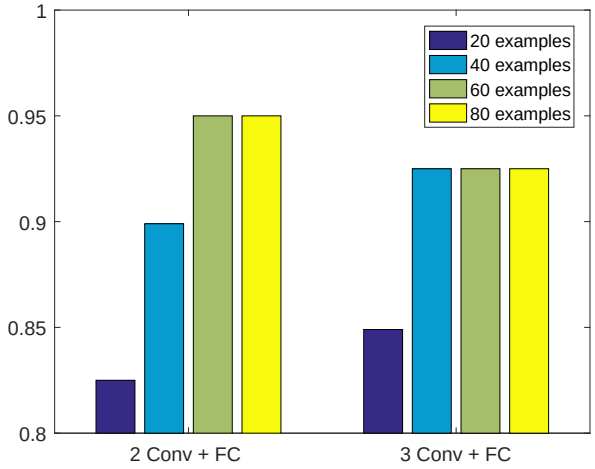

Figure 6. *Accuracy for a 2 Convoluational plus FC layer architecture.*

Figure 6 is in accordance to the previously reported findings. Indeed, deeper networks do not necessarily help in reaching better accuracy. More specifically, we observe that for 60 and 80 examples, the 2 layer approach offers better performance compared to the deeper three layer network. The phenomenon is attributed to the fact that the limited number of examples lead to overfitting of deeper networks. For limited number of training examples however, the three layer is able to attain a stable performance faster compared to the two layer case since there is a sufficient number of network parameters to capture the intrinsic characteristic of the data.

While in Figures 4,5 and 6, the network was trained with either pixel intensities or optical flow values, Figure 7 presents the accuracy achieved when both structural and motion features are fused and utilized for the final classification. The performance reported in Figure 7 clearly demonstrate the modeling capabilities achieved when both intensity and motion related features are extracted. Even for the two convolutional layer architecture, al-

though the performance is on par with the case of optical flow based approach, we observe that high classification accuracy is achieved from a small number of epochs.

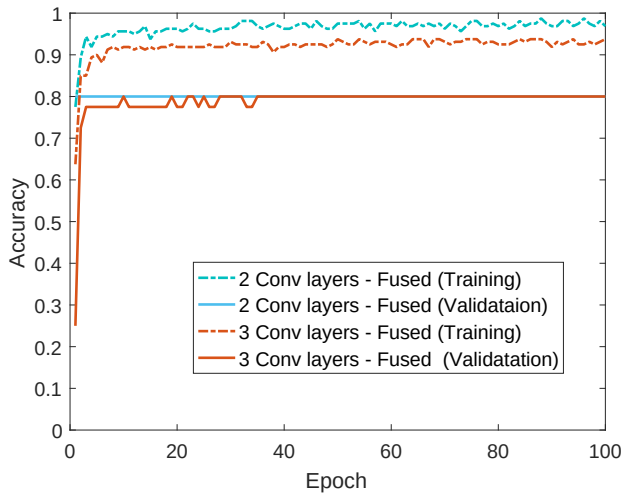

Figure 7. *Accuracy for fusion architecture.*

The last experimental results we report involve the use of pre-trained networks for extracting structural information from pixel intensities. More specifically, Figure 8 showcases the accuracy achieved using a two and a three layer CNN architecture applied of frame-level features extracted from a VGG16 network. The results suggest that using spatial only pre-trained models cannot lead to the performance attained when the higher order structure of the data is exploited.

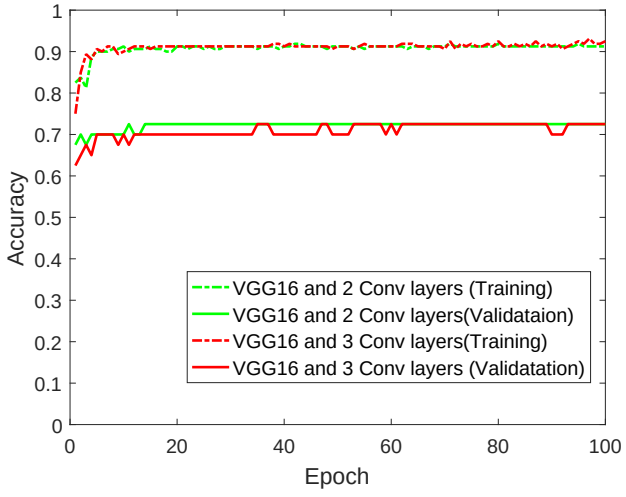

Figure 8. *Accuracy for VGG16 input data introduced to 2 Convoluational plus FC layer architecture.*

Conclusions

The accurate classification of video sequences requires addressing numerous theoretical and practical issues, related to the complexity and the size of the data. In this work, we consider the problem of automated identification of winning shot in broadcasted tennis games which is treated as a binary classification problem. We investigate the use of 3D Convolutional Neural Networks encoding both spatial and temporal information. Furthermore, we propose the fusion of the 3D CNN outputs for encoding both structural and motion related characteristics. Experimental results on a new manually annotated dataset demonstrate the ability of the proposed approach to infer contextual information.

Acknowledgments

This work was funded by the DEDALE project contract no.665044 within the H2020 Framework Program of the European Commission.

References

[1] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, pp. 1097-1105. 2012.

[3] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[4] K. Alex, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, pp. 1097-1105, 2012.

[5] Karalasa, Konstantinos, Grigorios Tsagkatakisb, Michalis Zervakisa, and Panagiotis Tsakalidesa. "Deep learning for multi-label land cover classification." In SPIE Remote Sensing, pp. 96430Q-96430Q. International Society for Optics and Photonics, 2015.

[6] Wang, Cong, Peng Zhang, Yanning Zhang, Lei Zhang, and Wei Wei. "A multi-label Hyperspectral image classification method with deep learning features." In Proceedings of the International Conference on Internet Multimedia Computing and Service, pp. 127-131. ACM, 2016.

[7] Li, Ke, Yalei Wu, Yu Nan, Pengfei Li, and Yang Li. "Hierarchical multi-class classification in multimodal spacecraft data using DNN and weighted support vector machine." Neurocomputing (2017).

[8] Abu-El-Haija, Sami, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. "YouTube-8M: A large-scale video classification benchmark." arXiv preprint arXiv:1609.08675 (2016).

[9] M. Jefferson Ryan, and A. Savakis. "Anomaly Detection in Video Using Predictive Convolutional Long Short-Term Memory Networks." arXiv preprint arXiv:1612.00390, 2016.

[10] Y.-H. Ng, J. M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. "Beyond short snippets: Deep networks for video classification." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4694-4702, 2015.

[11] Varol, G., Laptev, I. and Schmid, C., "Long-term temporal convolutions for action recognition." IEEE transactions on pattern analysis and machine intelligence. 2017.

[12] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolu-

tional networks for action recognition in videos." In Advances in neural information processing systems, pp. 568-576. 2014.

[13] Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks." In Proceedings of the IEEE international conference on computer vision, pp. 4489-4497. 2015.

[14] Diba, Ali, Ali Mohammad Pazandeh, and Luc Van Gool. "Efficient two-stream motion and appearance 3d cnns for video classification." arXiv preprint arXiv:1608.08851 (2016).

[15] P. Xiaojiang, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding, 2016.

[16] Ma, Shugao, Leonid Sigal, and Stan Sclaroff. "Learning activity progression in lstms for activity detection and early detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[17] B. Hakan, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. "Dynamic image networks for action recognition." In IEEE International Conference on Computer Vision and Pattern Recognition CVPR. 2016.

[18] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, CRCV-TR-12-01, November, 2012.

[19] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

Author Biography

Grigorios Tsagkatakis *received his B.S. and M.S. degrees in Electronics and Computer Engineering from Technical University of Crete, in 2005 and 2007 respectively. He was awarded his PhD in Imaging Science from the Center for Imaging Science at the Rochester Institute of Technology, USA in 2011. He is currently a postdoctoral fellow at the Institute of Computer Science - FORTH, Greece. His research interests include signal and image processing with applications in sensor networks and imaging systems.*

Mustafa Jaber *is a computer vision engineer at NantOmics in Culver City, California where he performs research and development work in the areas of deep learning and machine vision. Dr. Jaber received BS degree in electrical engineering from the Islamic University of Gaza, Gaza, Palestine, in 2003, and MS in the same discipline from the Rochester Institute of Technology (RIT), Rochester, New York, in 2007. Dr. Jaber also received a PhD in imaging science from RIT in 2012. His research interests are in the areas of digital image understanding, visual search, image ranking, sports vision, and medical image analysis.*

Panagiotis Tsakalides *received the Diploma degree from Aristotle University of Thessaloniki, Greece, and the PhD. degree from the University of Southern California, Los Angeles, USA, in 1990 and 1995, respectively, both in electrical engineering. He is a Professor and the Chairman with the Department of Computer Science, University of Crete, and Head of the Signal Processing Laboratory, Institute of Computer Science, Crete, Greece. He has coauthored over 150 technical publications, including 30 journal papers. He has been the Project Coordinator in seven European Commission and nine national projects. His research interests include statistical signal processing with emphasis in non-Gaussian estimation and detection theory, sparse representations, and applications in sensor networks, audio, imaging, and multimedia systems.*