Skeleton-based Action Recognition Based on Deep Learning and Grassmannian Pyramids

Dimitrios Konstantinidis, Kosmas Dimitropoulos and Petros Daras
ITI-CERTH, 6th km Harilaou-Thermi, 57001, Thessaloniki, Greece
Email:{dikonsta;dimitrop;daras}@iti.gr

***Abstract*—The accuracy of modern depth sensors, the robustness of skeletal data to illumination variations and the superb performance of deep learning techniques on several classification tasks have sparkled a renewed interest towards skeleton-based action recognition. In this paper, we propose a four-stream deep neural network based on two types of spatial skeletal features and their corresponding temporal representations extracted by the novel Grassmannian Pyramid Descriptor (GPD). The performance of the proposed action recognition methodology is further enhanced by the use of a meta-learner that takes advantage of the meta knowledge extracted from the processing of the different features. Experiments on several well-known action recognition datasets reveal that our proposed methodology outperforms a number of state-of-the-art skeleton-based action recognition methods.**

1. Introduction

Human action recognition has been a growing research area for the past decades due to its wide applicability to surveillance, video retrieval and human machine interaction. Traditionally, human action recognition has been achieved using RGB video sequences. However, the sensitivity of the RGB data to illumination changes, background clutter and occlusions and the technological advances in depth sensors has led to the introduction of skeletal data (i.e., set of joints in the 3D space) for human action recognition as they have proven to be robust to illumination variations, human scale and viewpoint. Although modern depth sensors can reliably extract 3D joint coordinates, skeleton-based action recognition remains a challenging problem due to variations in the way people perform actions and joint self-occlusions.

Below, we present the work that is most related to ours, however a comprehensive review of skeleton-based action recognition methodologies can be found in [1]. Except from raw joint coordinates, several recently proposed methods employ skeletal spatial and temporal features. Seidenari et al. in [2] proposed the decomposition of a skeleton in kinematic chains and the expression of joint coordinates in a chain in a local reference system, achieving rotation and translation invariance. On the other hand, Hussein et al. in [3] proposed a temporal covariance descriptor based on joint coordinates, while Wang et al. in [4] employed a kernel-based covariance matrix as a generic feature representation for skeleton-based action recognition. Pair-wise joint distances and joint differences between current and previous postures were employed in [5].

Zhou et al. in [6] proposed the extraction of discriminative action key poses based on normalized joint locations, velocities and accelerations, while Sharaf et al. in [7] employed a pyramid of covariance matrices to encode the relationship between joint angles and angular velocities. Xia et al. introduced Histograms of 3D Joint Locations by assigning joint positions into cone bins in 3D space [8]. Vemulapalli et al. proposed the representation of each skeleton sequence as a curve in the Lie group and achieved state-of-the-art performance in several action recognition datasets [9]. An actionlet ensemble model that captures local interactions in the form of relative spatial displacement between skeleton joints was proposed in [10]. Meshry et al. in [11] proposed gesturelets that encode the position and kinematic information of skeleton joints, while Patrona et al. in [12] extended gesturelets by adding automatic feature weighting at frame level and employing kinetic energy to identify the most representative action poses.

The outstanding performance of deep learning on several tasks has led to its use on skeleton-based action recognition as well. Zhang et al. proposed several geometric features that can be extracted from skeleton joints and fed them to a 3-layer Long Short-Term Memory (LSTM) network for accurate human action classification [13]. Taking a different approach, Wang et al. in [14] proposed Joint Trajectory Maps that compactly encode spatio-temporal information of 3D skeleton sequences into multiple 2D images.

Finally, based on linear dynamical systems (LDSs), which have been widely used in the past for dynamic texture classification [15], Dimitropoulos et al. proposed the representation of skeleton action sequences as clouds of points in a Grassmannian manifold [16]. Inspired by LDSs and the discriminative power of deep neural networks, we propose in this paper a novel method based on alternative temporal skeleton representations that can be combined in a deep learning framework to achieve state-of-the-art action recognition results. The main contributions of this work are: (a) a novel four-stream deep neural network that takes advantage of four different temporal skeleton representations to achieve accurate and robust action recognition results, (b) a novel descriptor (GPD) that captures dynamics of actions from different temporal levels and (c) the use of a meta-

ISBN 978-90-827970-1-5

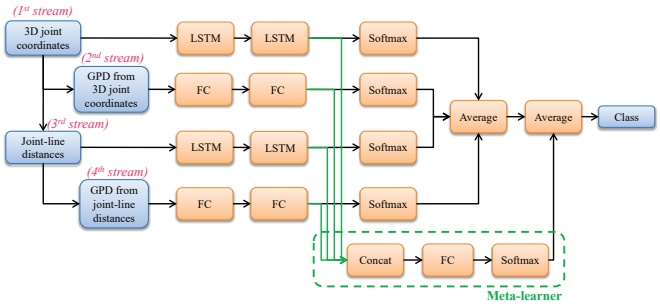

Figure 1. Proposed deep neural network architecture for skeleton-based action recognition.

learner, which is a network that exploits meta knowledge from the various streams to improve classification accuracy.

The remaining of this paper is organised as follows: Section 2 presents the proposed action recognition methodology, while Section 3 presents a comparative evaluation of our methodology with respect to other state-of-the-art algorithms on several action recognition datasets. Finally, Section 4 concludes the work of this paper.

2. Methodology

In order to accurately and reliably classify actions, we propose two skeletal spatial features. The first type of spatial features is the 3D joint coordinates that are computed based on a common preprocessing scheme, applied to the raw joint coordinates [9], [13]. More specifically, all 3D joint coordinates are initially transformed from the world to a person-centric coordinate system by placing the hip center at the origin. Afterwards, the body part lengths of all skeletons in a dataset are normalized (without changing joint angles) with respect to the corresponding lengths of a reference skeleton that is randomly chosen from the dataset. Finally, the skeletons are rotated in a way that the ground plane projection of the left to right hip vector is parallel to the global x-axis. Such a preprocessing makes skeletons invariant to the absolute location of the human in the scene, scale-invariant and view-invariant respectively.

The second type of spatial features that are employed in this work is the joint-line distances [13]. Joint-line distances model the distances from each joint to its projections on the lines formed by every other skeleton joint pair. Given three different joints of a skeleton $J_1, J_2, J_3 \in R^3$, the distance $d(J_1, L_{J_2 \to J_3})$ between joint J_1 and the line formed by J_2 and J_3, is given by employing Heron's formula as follows:

$$d(J_1, L_{J_2 \to J_3}) = \frac{2\sqrt{s(s - d(J_1, J_2))(s - d(J_2, J_3))(s - d(J_3, J_1))}}{d(J_2, J_3)} \quad (1)$$

,where $d(*, *)$ denotes the distance between two 3D joint coordinates and $s = 0.5(d(J_1, J_2) + d(J_2, J_3) + d(J_3, J_1))$. The motivation behind the selection of the joint-line distances lies in the fact that they consist an alternative spatial representation that models the relationship between skeleton joints. As a result, joint-line distances can complement 3D joint coordinates, forming a very descriptive representation that can significantly improve action recognition results.

Based on the previously defined spatial features, we propose a four-stream deep neural network, as shown in Figure 1. Both 3D joint coordinates and joint-line distances follow a two-stream processing stage. In the 1st and 3rd streams, the spatial features are directly fed to LSTMs in order to derive temporal information, while in the 2nd and 4th streams, novel GPD features, described in detail in Section 2.1, are extracted from the spatial features and model the temporal dynamics of these multi-dimensional signals. Afterwards, the GPD features are processed with fully connected layers. The processed features are then fed to softmax classifiers in order to derive probabilities for each class, before these probabilities are fused (i.e., averaged) to get an overall prediction. The motivation behind the proposed network is the construction of four different temporal representations of the same skeleton sequence. Given that a single temporal representation may not be descriptive enough for each tested dataset, the use of four complementary temporal representations within a deep network that can weigh them accordingly can assist in improving action recognition results both in the same and across different datasets.

Finally, a meta-learner, described more extensively in Section 2.2, is employed that concatenates the temporal features computed from the four streams of the proposed deep network and processes them in order to derive even more discriminative features. These features are then fed to another softmax classifier before the predictions from the meta-learner and the fusion of the four streams are averaged to give the final prediction.

2.1. LDS theory and Grassmannian pyramids

According to the LDS theory, the stochastic modeling of both signal dynamics (represented as a time-evolving hidden state process $x(t) \in R^n$) and appearance ($y(t) \in R^d$, where d is the length of the input signal per frame) is encoded by the following two stochastic processes:

$$\begin{aligned} x(t+1) &= Ax(t) + Bv(t) \\ y(t) &= \bar{y} + Cx(t) + w(t) \end{aligned} \quad (2)$$

,where $A \in R^{n \times n}$ is the hidden state transition matrix, while $C \in R^{d \times n}$ represents the mapping of the hidden state to the output of the system. The quantities $w(t) \sim N(0, R)$ and $Bv(t) \sim N(0, Q)$ are the measurement and process noise respectively, while $\bar{y} \in R^d$ is the mean value of the observed data. The LDS descriptor, $M_{LDS} = (A, C)$, contains both the appearance information of the observed data modeled by C, and its dynamics that are represented by A. Dimitropoulos et al. [16] proposed a higher-order LDS, where a temporal sequence is split in segments, the LDS descriptor of each segment is mapped to a point in the Grassmannian manifold and these points are then clustered to form a Histogram of Grassmannian points (HoGP).

ISBN 978-90-827970-1-5

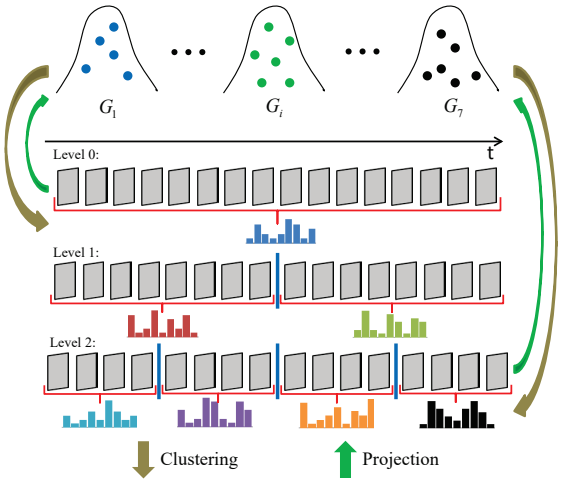

Figure 2. A temporal sequence is split in segments, LDS descriptors are projected to Grassmannian manifolds and histograms of Grassmannian points are extracted and then concatenated to form the final GPD.

In this work, we extend HoGP descriptors and propose the novel Grassmannian Pyramid Descriptor (GPD). The motivation behind GPDs is the construction of a temporal representation with the ability to capture dynamics of a multi-dimensional signal (i.e., 3D joint coordinates and joint-line distances in this case) in different temporal resolutions and of different segments. A GPD consists of three levels, where in each subsequent level both the temporal sequence and the window size that splits the sequence in segments are halved (see Figure 2). As a result, a temporal sequence can now be represented both in coarser levels, achieving robustness to noise, and in finer levels, paying more attention to details. Moreover, the proposed GPD representation can effectively handle temporal scale variations.

Additionally, the LDS descriptors, extracted from each temporal segment are projected as Grassmannian points to different manifolds. As a result, the proposed GPD representation leads to seven Grassmannian manifolds G_i, $i = \{1, ..., 7\}$, where each Grassmannian manifold describes the dynamics from specific segments of the temporal sequences and of different temporal resolutions. The clustering of Grassmannian points is performed on each manifold separately and thus a histogram is computed for each manifold. Finally, the histograms are concatenated into a larger histogram that consists the novel GPD representation. As we show in the experimental section of this paper, such a representation enhances the discrimination ability of the proposed methodology in the task of action recognition.

2.2. Exploiting meta knowledge

Meta-learning is a sub-field of machine learning, where an automatic algorithm is applied on the meta data of different classifiers in order to improve their combined classification accuracy. Recently, meta-learners have been employed in deep learning architectures as well [17]. In this work, we propose a meta-learner (see dotted outline in Figure 1) that is applied on the features computed from the streams of our deep model instead of being applied on their corresponding predictions (i.e., after the softmax classifiers) and fuses its prediction with the average prediction of the softmax classifiers, significantly differentiating from previous uses of a meta-learner in a deep learning framework.

The motivation behind the use of a meta-learner is the fact that a classifier introduces inductive bias, meaning that the classifier's assumptions about a problem and the data can make it effective only on similar types of problems. Although this work deals specifically with skeleton-based action recognition, the variations in skeleton acquisition procedures, number of joints and types of actions introduced by the different datasets can significantly affect classifiers rendering them unable to perform optimally across all datasets. Furthermore, features should be weighted differently when applied on different datasets as their contribution to the action recognition task usually varies depending on the current set of actions that needs to be identified.

The proposed meta-learner combines the features learned from the streams of our deep model appropriately in order to come up with even more discriminative features. The output of the meta-learner is finally fused (i.e averaged) with the average output of our deep model. Thus, the proposed meta-learner is integrated in our deep model, assisting in its optimization during the training phase. In this way, we enhance the learning procedure and improve the discrimination and generalization ability of the proposed action recognition methodology.

3. Experimental evaluation

3.1. Datasets and evaluation settings

Four datasets are employed for the evaluation of our methodology as shown below. The selection of these datasets is based on their different characteristics (i.e types of actions, number of joints, etc.) that pose challenges to a general action recognition method. Furthermore, the small size of two of these datasets introduces difficulties to the training of a deep neural network that usually requires an abundance of training samples.

UT-Kinect dataset [8]: This dataset consists of 10 actions performed twice by 10 different subjects. Each skeleton consists of 20 joints. For the evaluation of this dataset, we follow the cross-subject test setting of [9], in which 10 folds are created, where half of the subjects are used for training and the remaining half for testing.

Florence3D Actions dataset [2]: This dataset consists of 9 actions performed two or three times by 10 different subjects. Each skeleton consists of 15 joints. For the dataset evaluation, we follow the cross-subject test setting of [9].

G3D Gaming Action dataset [18]: This dataset consists of 20 actions performed three times by 10 different subjects. Each skeleton consists of 20 joints. For the evaluation of this dataset, we follow the protocol of [16], in which the first instance of each action per subject is used for training and the other two instances are used for testing.

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE 1. CLASSIFICATION ACCURACY ON MSRC-12 USING (A) CROSS-SUBJECT PROTOCOL [14] AND (B) MODALITY-BASED "LEAVE-PERSONS-OUT" PROTOCOL [12].

Method	Accuracy
ConvNet+JTM [14]	93.12%
Ker-RP [4]	92.3%
Cov3DJ [3]	91.7%
ELC-KSVD [6]	90.22%
Proposed	**94.65%**

(a)

Modality \ Method	Sharaf et al. [7]	Meshry et al. [11]	Patrona et al. [12]	Proposed
Video	0.669 ± 0.082	0.895 ± 0.068	0.927 ± 0.009	**0.969 ± 0.069**
Image	0.598 ± 0.082	0.858 ± 0.086	0.894 ± 0.010	**0.944 ± 0.091**
Text	0.558 ± 0.092	0.788 ± 0.139	0.851 ± 0.012	**0.871 ± 0.165**
Video-Text	0.684 ± 0.074	0.921 ± 0.126	0.983 ± 0.008	**0.992 ± 0.024**
Image-Text	0.687 ± 0.099	0.894 ± 0.085	0.905 ± 0.007	**0.956 ± 0.089**
Overall	0.639	0.871	0.912	**0.946**

(b)

MSRC-12 Kinect Gesture dataset [19]: This large dataset consists of 30 actions performed by 12 subjects. Each skeleton consists of 20 joints. For the dataset evaluation, we follow two protocols; a cross-subject protocol [14], where the odd subjects are used for training and the even subjects for testing and a modality-based "leave-persons-out" protocol [12], in which all but one subjects are used for training and the remaining subject for testing for each modality (i.e., video, image, text, video-text, image-text). The ground truth annotation of this dataset is based on [3].

The skeleton sequences of all datasets are processed so that they are composed of 64 frames either by removing intermediate frames in the case of larger sequences or by adding interpolated intermediate frames in the case of smaller sequences.

3.2. Model parameters

The parameters that affect our proposed methodology (i.e., size of layers, dropout, learning rate, etc.) are determined after experimentation on the UT-Kinect dataset and kept fixed for the other datasets. In this way, we want to point out the advantages of the proposed features and the meta-learner on the performance of our methodology, no matter which dataset is employed. More specifically, the two-layer LSTMs consist of 1024 and 256 neurons and dropout/recurrent dropout equal to 0.1 and 0.2 respectively. Furthermore, the fully connected layers (FCs) that are fed with the GPDs consist of 512 and 128 neurons respectively, while the FC layer of the meta-learner consists of 128 neurons. The window size for the computation of GPDs is halved in each subsequent level from 16 to 4 frames. Finally, the network is implemented in Keras-Tensorflow framework and trained using the Adam optimizer with batch size of 32 and learning rate equal to 0.0001.

3.3. Results

In this section, our proposed methodology is compared with 15 state-of-the-art action recognition methods across four datasets. Table 1 evaluates the performance of our methodology on the large MSRC-12 dataset. It can be observed that our method outperforms all other state-of-the-art methods in both evaluation settings and in all modalities, achieving a significant boost on the accuracy. More specifically, our methodology improves the state-of-the-art results by 1.53% and 3.4% when the cross-subject and modality-based "leave-persons-out" protocol are employed respectively.

TABLE 2. CLASSIFICATION ACCURACY ON G3D DATASET. ALL METHODS WERE TAKEN FROM [16].

Method	Accuracy
Sh-LDS-HoGP [16]	90.75%
Restricted Boltzmann Machine	84%
Hidden Markov Model	77.4%
Conditional Random Fields	69.25%
Dynamic Time Warping	57%
Proposed	**92.38%**

TABLE 3. CLASSIFICATION ACCURACY ON UT-KINECT DATASET.

Method	Accuracy
Lie Group [9]	97.08%
Histogram of 3D joints [8]	90.92%
Random forests [5]	87.9%
Proposed	**97.69%**

TABLE 4. CLASSIFICATION ACCURACY ON FLORENCE3D DATASET.

Method	Accuracy
Lie Group [9]	90.88%
Multi-Part Bag-of-Poses [2]	82.00%
Proposed	**91.12%**

TABLE 5. EXPERIMENTATION WITH PROPOSED CONTRIBUTIONS ON UT-KINECT DATASET.

Contributions	Accuracy
HoGP [16] from 3D joint coordinates	68.55%
GPD from 3D joint coordinates	**81.31%**
HoGP [16] from joint-line distances	60.60%
GPD from joint-line distances	**78.80%**
Proposed without meta-learner	96.38%
Proposed with meta-learner	**97.69%**

Similar performance improvement is noticed for the other tested datasets as well, although their small sizes poses challenges to the accurate training of our proposed deep neural network. From Tables 3 and 4, it can be observed that our methodology outperforms the Lie Group method by 0.61% and 0.24% on the UT-Kinect and Florence3D datasets respectively. Moreover, Table 2 shows that our proposed methodology outperforms the Sh-LDS-HoGP method and other classification approaches by at least 1.63%, meaning that the proposed features are more descriptive of the under-

ISBN 978-90-827970-1-5 © EURASIP 2018

lying actions of the G3D dataset than the HoGP features.

The superb performance of the proposed methodology across all tested datasets reveals the splendid ability of the meta-learner to weigh the different features in a way that makes our method achieve similar performance irrespective of the tested dataset. At this point, it is worth noting that the hyper-parameters of the proposed deep network are kept fixed after their optimization with respect to the UT-Kinect dataset. As a result, we can conclude that the proposed methodology generalizes well on other datasets without requiring additional hyper-parameter tuning.

Finally, we analyse the effect of our contributions on the classification accuracy of our proposed deep model on the UT-Kinect dataset. From studying Table 5, we can observe the huge boost on the classification accuracy of the proposed deep network when the novel GPDs are employed. More specifically, an improvement of 18.6% is observed when the HoGP features are substituted with the GPDs extracted from the 3D joint coordinates. A similar improvement is noticed in the case of GPDs extracted from joint-line distances. This means that the proposed GPDs are successful in their task of enhancing the discrimination ability of the proposed deep network. Finally, the introduction of the meta-learner in the proposed deep network leads to a better exploitation of the meta knowledge derived from the four network streams and improves the classification accuracy of the proposed methodology by almost 1.35%.

4. Conclusions

A novel skeleton-based action recognition method that employs joint coordinates, joint-line distances and their temporal representations through the use of the novel GPDs is proposed in this work. A meta-learner that appropriately combines the meta knowledge derived from the four streams of the proposed deep network is also employed. Experimentation on several well-known action recognition datasets reveals that the proposed methodology achieves state-of-the-art performance and demonstrates improved discrimination ability across datasets with different sets of actions.

Acknowledgments

This work has been supported from EC under grant agreement no. H2020-ICT-19-2016-2 “EasyTV”.

References

[1] F. Han, B. Reily, W. Hoff, and H. Zhang, “Space-time representation of people based on 3D skeletal data: A review,” *Computer Vision and Image Understanding*, vol. 158, pp. 85–105, 2017.

[2] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, “Recognizing Actions from Depth Cameras as Weakly Aligned Multi-part Bag-of-Poses,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 479–485.

[3] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, “Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations,” in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, 2013, pp. 2466–2472.

[4] L. Wang, J. Zhang, L. Zhou, C. Tang, and W. Li, “Beyond Covariance: Feature Representation with Nonlinear Kernel Matrices,” in *IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 4570–4578.

[5] Y. Zhu, W. Chen, and G. Guo, “Fusing Spatiotemporal Features and Joints for 3D Action Recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 486–491.

[6] L. Zhou, W. Li, Y. Zhang, P. Ogunbona, D. T. Nguyen, and H. Zhang, “Discriminative Key Pose Extraction Using Extended LC-KSVD for Action Recognition,” in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov 2014, pp. 1–8.

[7] A. Sharaf, M. Torki, M. E. Hussein, and M. El-Saban, “Real-Time Multi-scale Action Detection from 3D Skeleton Data,” in *IEEE Winter Conference on Applications of Computer Vision*, Jan 2015, pp. 998–1005.

[8] L. Xia, C. Chen, and J. Aggarwal, “View invariant human action recognition using histograms of 3D joints,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 20–27.

[9] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 588–595.

[10] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Learning Actionlet Ensemble for 3D Human Action Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914–927, May 2014.

[11] M. Meshry, M. E. Hussein, and M. Torki, “Linear-time online action detec- tion from 3D skeletal data using bags of gesturelets,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan 2016, pp. 1–9.

[12] F. Patrona, A. Chatzitofis, D. Zarpalas, and P. Daras, “Motion analysis: Action detection, recognition and evaluation based on motion capture data,” *Pattern Recognition*, vol. 76, pp. 612–622, 2018.

[13] S. Zhang, X. Liu, and J. Xiao, “On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2017, pp. 148–157.

[14] P. Wang, Z. Li, Y. Hou, and W. Li, “Action Recognition Based on Joint Trajectory Maps Using Convolutional Neural Networks,” in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 102–106.

[15] K. Dimitropoulos, P. Barmpoutis, and N. Grammalidis, “Higher Order Linear Dynamical Systems for Smoke Detection in Video Surveillance Applications,” *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 27, no. 5, pp. 1143–1154, 2017.

[16] K. Dimitropoulos, P. Barmpoutis, A. Kitsikidis, and N. Grammalidis, “Classification of Multidimensional Time-Evolving Data using Histograms of Grassmannian Points,” *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 28, no. 4, pp. 892–905, 2018.

[17] S. Amiri, M. Pourazad, P. Nasiopoulos, and V. Leung, “Human action recognition using meta learning for RGB and depth information,” in *International Conference on Computing, Networking and Communications (ICNC)*, Feb 2014, pp. 363–367.

[18] V. Bloom, D. Makris, and V. Argyriou, “G3D: A gaming action dataset and real time action recognition evaluation framework,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2012, pp. 7–12.

[19] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, “Instructing People for Training Gestural Interactive Systems,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1737–1746.

ISBN 978-90-827970-1-5 © EURASIP 2018