

# A compact sequence encoding scheme for online human activity recognition in HRI applications

Georgios Tsatiris<sup>1</sup>, Kostas Karpouzis<sup>1</sup>, and Stefanos Kollias<sup>1,2</sup>

<sup>1</sup> Artificial Intelligence and Learning Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, 9 Heroon Politechniou str., 15780, Athens, Greece

<sup>2</sup> School of Computer Science, University of Lincoln, Brayford Pool, Lincoln, Lincolnshire, UK  
gtsatiris@image.ntua.gr, kkar pou@cs.ntua.gr, stefanos@cs.ntua.gr

**Abstract.** Human activity recognition and analysis has always been one of the most active areas of pattern recognition and machine intelligence, with applications in various fields, including but not limited to exertion games, surveillance, sports analytics and healthcare. Especially in Human-Robot Interaction, human activity understanding plays a crucial role as household robotic assistants are a trend of the near future. However, state-of-the-art infrastructures that can support complex machine intelligence tasks are not always available, and may not be for the average consumer, as robotic hardware is expensive. In this paper we propose a novel action sequence encoding scheme which efficiently transforms spatio-temporal action sequences into compact representations, using Mahalanobis distance-based shape features and the Radon transform. This representation can be used as input for a lightweight convolutional neural network. Experiments show that the proposed pipeline, when based on state-of-the-art human pose estimation techniques, can provide a robust end-to-end online action recognition scheme, deployable on hardware lacking extreme computing capabilities.

**Keywords:** Action recognition · Deep neural networks · Human-Robot Interaction · Radon transform.

## 1 Introduction

Human activity recognition is a very important field of study inside the broad domain of pattern recognition, with many everyday life applications. High level human actions are an important information modality for social interaction and they play a critical role in Human-Robot Interaction (HRI) [12]. Human poses, hand gestures and overall body motion are also important for human-focused HRI, especially in applications with robots interacting with humans daily (e.g. [18], [1] and [20]). Robotic assistants are required to understand and analyse human activity at all levels of interpretation, as well as have the ability to predict intentions and to imitate human actions when forming responses. Social robots

and applications aimed at such environments will need to be employed in human-centric contexts, build trust and effectively assist humans beings [6]. In this paper, we focus on online human activity recognition pipelines with the aim to provide solutions on the inherent problems of such endeavours.

### 1.1 Related Work

Various works on activity analysis and understanding deal with real application scenarios. In the field of preventing unintentional falls in elderly care and assistive environments, the work presented in [29] is one of the most typical attempts for applying convolutional neural networks in a fall detection application. A dataset of daily activities recorded from many people is used as classification is performed on single frames containing human postures. This method performs well on posture classification and can detect falls with a low false alarm rate. In an earlier work [16], an efficient algorithm is proposed to identify temporal patterns in actions. It further utilizes these patterns to construct activity representations for automated recognition.

There is an extensive literature of studies focused on HRI, a comprehensive presentation of which can be found in [12]. In a typical example of such works, authors in [22] propose that state-of-the-art approaches from automatic speech and visual action recognition, fused with multiple modalities, can construct multimodal recognition pipelines focused on assistive robots for elderly environments. An end-to-end approach shown in [3] deals with dynamic gestures using a Dynamic Time Warping approach, based on features extracted from depth data, in a complete HRI scenario.

### 1.2 Real world challenges

As stated in [25], applied human activity recognition pipelines are demanding in terms of hardware and general infrastructure. Most action recognition systems, especially smart-home or assistive living applications, depend on network infrastructures for easy data fusion and integration of different sensing modalities. HRI environments are no different, in the sense that they may need computing capabilities similar to common workstations, as well as networking. Especially when dealing with deep learning-based pipelines, we are used to having hardware acceleration (i.e. massive parallel computing, GPU acceleration etc.) at our disposal. This study aims to provide a lightweight pipeline with limited need for high-end infrastructures.

## 2 Theoretical Primer

In the following section, we attempt a brief documentation of the theoretical background on which the proposed scheme is built upon. This primer emphasizes on the statistical nature of spatio-temporal data, the use of the Radon Transform for feature encoding, as well as the necessary tools to encode the necessary information into a compact pipeline.

## 2.1 Variance-based Features

As mentioned in [24], high level human pose and activity descriptors, such as spatio-temporal interest points [14] or human joints, can be treated as probabilistic distributions. Joints detected in a video frame or still image, e.g. using pipelines such as OpenPose [4], can be considered distributed in space, whereas spatio-temporal interest points are distributed both in space and time. Action recognition techniques benefit from accurately modeling these relationships between the salient features describing an action sequence.

Particularly in [24], a method to distinguish between experienced tennis players and amateurs is shown, which employs feature vectors constructed by calculating the per frame variance of interest points. Variance, as the second central moment, is an effective shape descriptor [17]. Interest points and joint locations form a 3D volumes in space and time. In that context, we extend the idea demonstrated in [24], by incorporating the use of Mahalanobis distances.

**Mahalanobis Distance** Given a sample vector  $\mathbf{x}$  from a distribution  $D$  with mean  $\boldsymbol{\mu}$  and covariance matrix  $S$ , the Mahalanobis distance between  $\mathbf{x}$  and  $D$  is given by equation 1.

$$mahal(\mathbf{x}, \boldsymbol{\mu}, S) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T S^{-1} (\mathbf{x} - \boldsymbol{\mu})} \quad (1)$$

In this work, we are focusing on detected human joints in 3D space, like the ones extracted using depth sensors (e.g. Intel RealSense<sup>3</sup> or Microsoft Azure Kinect<sup>4</sup>) and by pipelines such as the one presented in [19], which extends common 2D heatmap-based pose detection techniques such as [4] without utilizing costly volumetric operations. There is a good indication that skeletal joints can be considered valid spatio-temporal interest points when it comes to human action analysis and, with them being a more refined and targeted source of human posture information, can lead to better results [27].

As the relative position of a joint and the human body is crucial in the definition of human motion, we believe that a compact and robust way to model these individual relations should be central to any action recognition pipeline. The Mahalanobis distance, as an established metric for point-to-distribution distances, is more suitable in this task, instead of more intuitive metrics, such the Euclidean distance between a point and the distribution centroid. Its place in this pipeline will be explained in detail in subsection 3.1.

## 2.2 Use of the Radon Transform

The Radon transform is effectively used as a reconstruction and feature extraction tool for several decades [2]. It has found applications in earlier human activity recognition pipelines based on hand-crafted features [7] and in novel techniques, alongside its generalization, the Trace transform [13], in offline pipelines

<sup>3</sup> <https://www.intelrealsense.com/>

<sup>4</sup> <https://azure.microsoft.com/en-us/services/kinect-dk/>

[10][9]. Another application of the Radon transform in this field leverages the transform’s ability to create translation and scale invariant features to use as input for a Recurrent Neural Network (RNN) [26]. However, for reasons which will be demonstrated in the next section, this paper avoids the use of RNN and draws inspiration from such studies as the one in [11], where source images and their Conic Radon transforms are fed into a deep convolutional neural network for fingerprint classification.

Given a 2D function  $f(x, y)$  with  $x, y \in \mathbb{R}$ , its Radon transform  $R_f$  is equal to the integral of all values of  $f$  under all lines  $L$ , defined by parameters  $\rho, \theta \in \mathbb{R}$ , where  $\rho$  is each line’s distance from the origin and  $\theta$  is the angle formed between the line and the horizontal axis:

$$R_f(\rho, \theta) = \int_L f(x, y) dL \quad (2)$$

If we substitute the parametrical  $\rho, \theta$  line equation in equation 2, we have the following:

$$R_f(\rho, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \delta(x \cos \theta + y \sin \theta - \rho) dx dy \quad (3)$$

where  $\delta$  the Dirac delta function, ensuring that only the values of  $f$  under line  $\rho, \theta$  will be integrated over.

### 3 The Proposed Pipeline

At this point, we have documented the theoretical tools to give spatio-temporal prominence to raw joint information using variance-based shape features and, particularly, the Mahalanobis distance. We have also seen how the Radon transform and other relevant transforms can create robust features out of otherwise non-salient 2D and 3D functions. This section will clarify how the aforementioned theoretical tools will be used in a unified and compact pipeline action encoding pipeline.

#### 3.1 Spatio-temporal Radon Footprints

As we have established, the relationship between the position of one joint and the rest of the body parts of a subject, at a particular point in time, can be encoded by treating the set of joints as a distribution and calculating the Mahalanobis distance (given by equation 1) between it and that particular joint. In our context, a point in time is a frame in an action sequence or video. So, if we extend the above claim to all the joints of a subject and across all frames of an action sequence, we can formulate a Mahalanobis matrix  $M$ , whose values for joint  $j$  and frame  $f$  can be calculated using the following equation:

$$M(j, f) = mahal(\mathbf{x}_f^j, \boldsymbol{\mu}_f, S_f) \quad (4)$$

where  $\mathbf{x}_f^j$  is the coordinate vector of joint  $j$  in frame  $f$ , defined in  $\mathbb{R}^2$  or  $\mathbb{R}^3$ ,  $\boldsymbol{\mu}_f$  is the mean of all joint coordinate vectors in frame  $f$  (essentially the centroid) and  $S_f$  is the covariance matrix of the joint distribution at frame  $f$ .

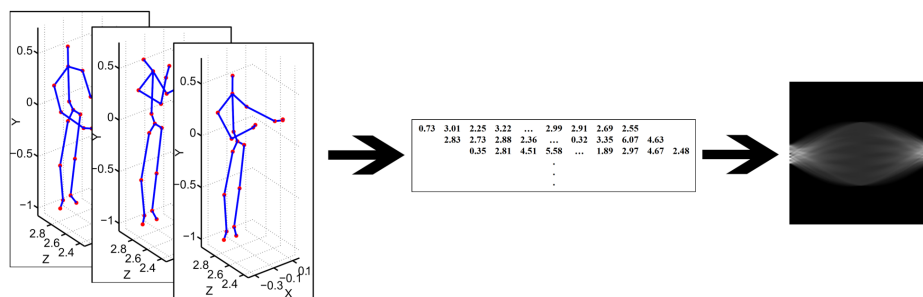
In essence, the Mahalanobis matrix formulated in equation 4 is a complete 2D representation of the action sequence in space and time. With  $J \times t$  resolution, where  $J$  is the total number of joints and  $t$  is the point in time (number of frames passed) since the beginning of the sequence, each line of the matrix has the Mahalanobis distances of every joint at the corresponding frame.

What this representation lacks, however, is a way to correlate the individual frame information into a compact robust feature which will describe the action sequence from its start up to point  $t$ . For this reason, we calculate the Radon transform of the 2D function  $M$  by substituting the representation  $M(j, f)$  into equation 3. This gives as a 2D spatio-temporal feature describing the action sequence up to point  $t$  in time. This feature will, for the rest of this paper, be called the Spatio-temporal Radon Footprint of the action up to point  $t$ , will be denoted as  $SRF_t$  and is calculated as follows:

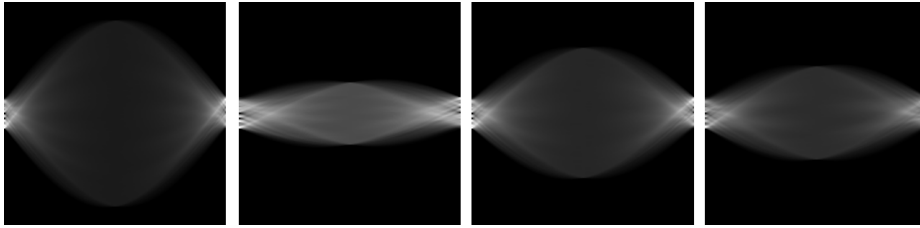
$$SRF_t(\rho, \theta) = \int_0^{t-1} \int_0^{J-1} M(j, f) \delta(j \cos \theta + f \sin \theta - \rho) dj df \quad (5)$$

In the above equation, it must hold that  $t \geq 2$ , both for  $M$  to be a 2D function of  $j$  and  $f$  and for the representation to encode a minimum of temporal information. In practice, depending on the action itself, more frames may need to pass before starting to calculate  $SRFs$ . Naturally, it should also hold that  $F > 2$ , which is obvious, both for the aforementioned reason as well as because it is impossible to ascertain the nature of an activity with insufficient joint information.

The process of calculating a  $SRF$  at a point in time, from a sequence of frames containing joint information, is depicted in figure 1. The frames shown are samples from the UTD-MHAD dataset [5], which will be documented in subsection 4.1. Figure 2 shows sample  $SRFs$  taken from this dataset, from different sequences and at different timestamps.



**Fig. 1.** Calculating  $SRF_t$ : The Mahalanobis matrix is calculated from the action sequence at point  $t$  and the Radon transform of the matrix gives the final result.



**Fig. 2.** *SRF* samples from various action sequences of the UTD-MHAD dataset, at varying points in time.

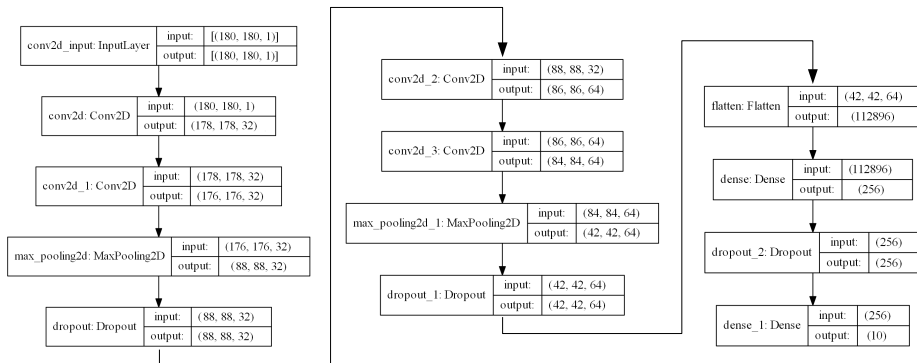
### 3.2 A VGG CNN-based pipeline

In previous works such as the ones in [10] and [9], the Radon transform and its derivatives were intermediate features, used to produce the final feature vector describing the action sequence, which in turn was used as input for established machine learning techniques, such as SVMs. In this work, however, much like in [11], instead of hand-crafting features based on the transform’s result, a convolutional neural network will learn those features directly from the *SRFs*. Particularly, we opted for a VGG-based architecture [23], because of the simplicity, the ease to train and deploy and the ability to adapt to complex pattern recognition problems that this family of convolutional neural networks demonstrates. The network used in our pipeline is shown in figure 3.

**Preference over RNN-LSTM architectures** It is customary to use Recurrent Neural Network (RNN, LSTM) architectures for online pattern recognition tasks involving temporal sequences [15], due to their ability to learn and encode temporal information. However, known issues come with the use of such networks, including the need for large training datasets in most cases, the vanishing and exploding gradient problems, as well their high demands in computational resources [8]. By encoding online spatio-temporal information in *SRFs* and using robust object detection architectures, we aim to eliminate the need for such networks and provide an online and compact action recognition pipeline which would not depend on extremely sophisticated hardware.

**Per-frame and cumulative classification** Human activity recognition is a temporally sensitive task. Offline techniques need to observe the complete action sequence before performing classification. In an online pipeline aiming towards real-time action recognition, per-frame classification is central. Previous studies, such as the one in [29], perform classification between unintentional falls and other activities on single postures/frames, rather than complete sequences. The methodology presented in this paper performs online action analysis, while taking into account the history of every previous per-frame classification task.

In a specific frame  $f$  of the sequence, the  $SRF_f$  is calculated and fed in the aforementioned network. The network then determines the class of the action



**Fig. 3.** The VGG-based architecture used in our experiments.

at that frame. However, the final decision is made by counting the number each action class was deemed the correct class of the sequence, divided by the number of frames which have passed up until the moment of classification. In a sense, for each action class and in every frame, we calculate a confidence that this action belongs to that class. This confidence is updated per-frame, with the ultimate goal of determining the correct class as early in the sequence as possible. This way, we eliminate the possibility of outliers (frames being classified in the wrong class) among a number of correct classifications.

## 4 Experimental Evaluation

In this section, we will discuss the experimental setup and the data used in the evaluation of the aforementioned methodology. The choice of the validation protocol will also become apparent, as well as the efficiency of the proposed technique.

### 4.1 Dataset

In our experiments, we used the UTD-MHAD dataset [5]. In particular, we used a more recent subset of this dataset, which contains data captured using the Kinect v2 depth sensor. It includes 10 actions, namely "Right hand high wave", "Right hand catch", "Right hand high throw", "Right hand draw X", "Right hand draw tick", "Right hand draw circle", "Right hand horizontal wave", "Right hand forward punch", "Right hand hammer" and "Hand clap (two hands)". These actions were performed by 6 subjects (3 female and 3 male), each one of

whom performed each action 5 times, resulting in a total of 300 action sequences. Figure 4 shows sample data from the dataset.

We opted for this new subset of the UTD-MHAD dataset for two reasons, both of which are sensor related. The Kinect V2 is a relatively state of the art device which can detect joint information for 25 body joints without inducing much noise (Base of the spine, Middle of the spine, Neck, Head, Left shoulder, Left elbow, Left wrist, Left hand, Right shoulder, Right elbow, Right wrist, Right hand, Left hip, Left knee, Left ankle, Left foot, Right hip, Right knee, Right ankle, Right foot, Spine at the shoulder, Tip of the left hand, Left thumb, Tip of the right hand, Right thumb).

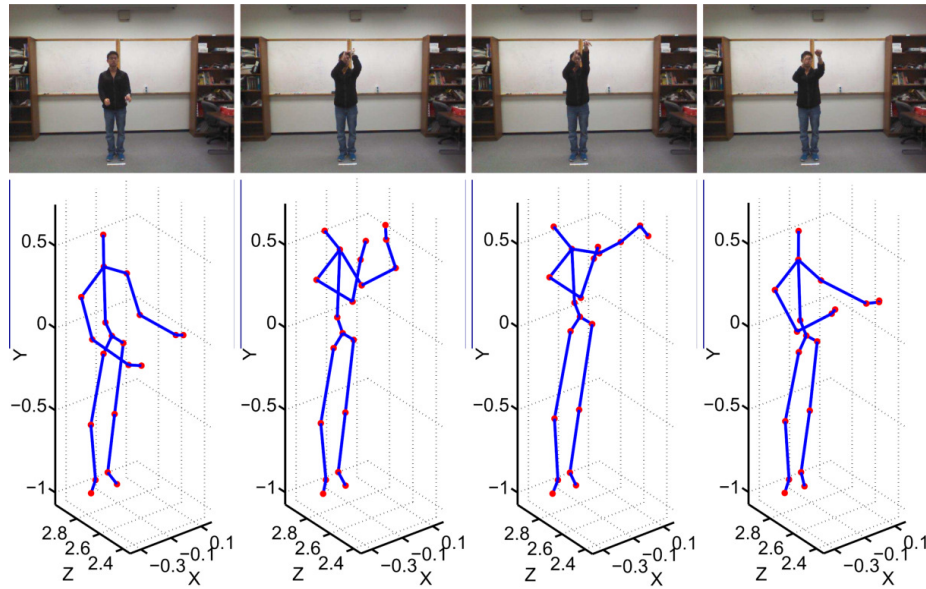


Fig. 4. Samples from the UTD-MHAD dataset.

## 4.2 Leave-One-Person-Out Cross-validation

Validating a human activity recognition pipeline entails an inherent complexity which is not found in other classification tasks. In a typical scenario, we would have partitioned the dataset into a training and a testing set. Our dataset, as many other action datasets, is already partitioned by the subjects executing the actions. Simply leaving a person out for validation and training with the rest would not, however, produce reliable results.

Every person performs an action in a unique to every other person’s way. For that matter, we perform the Leave-one-person-out cross-validation protocol (LOPO or LOO) across all individuals in our dataset. Essentially, we are training



with the actions performed by 5 subjects and test with the actions performed by the one we left out. We repeat this process for all subjects in the dataset. The final reported accuracy would be the mean of the individual accuracies of each training-testing scenario.

### 4.3 Results

One of the first findings that we should report is that our pipeline achieved high training accuracy across all individual subject training tasks, with the mean training accuracy reaching **99.15%**. The highest achieved training accuracy was as high as **99.25%**, when training with all subjects except for subject #3, which was used for testing.

An expected discrepancy was noticed across the validation tasks. Notably, when testing with subject #6, the pipeline achieved validation accuracy equal to **90.08%**, whereas the worst performance was reported when testing with with subject #3 (**66.02%**), on the same task that achieved the highest training accuracy. This hints towards possible issues of overfitting which may need to be dealt with in future works. The mean validation accuracy achieved after performing the complete LOPO protocol across all subjects was equal to **81.06%**. Table 1 shows the confusion matrix of the proposed pipeline after the validation process.

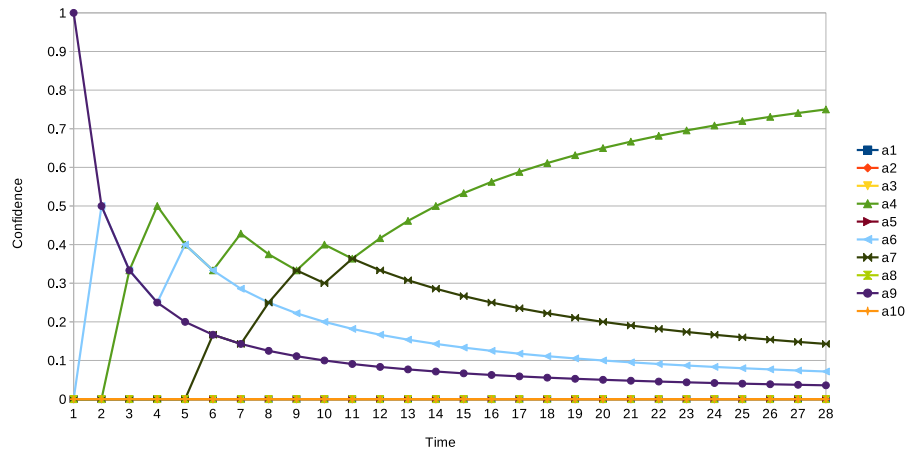
Figure 5 shows a characteristic sample of the evolution of class confidences according to our proposed pipeline. In this figure we observe how action class 4 ("Right hand draw X") emerges as the most possible to be the correct one with the passage of time. More extensive experimentation may lead to more robust findings concerning the similarity and relationship between classes.

**Table 1.** Confusion matrix of the proposed pipeline.

		Truth										
		a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	Total
Predicted	a1	<b>20</b>	0	0	0	0	0	2	0	1	0	23
	a2	0	<b>10</b>	1	0	0	0	0	0	0	0	11
	a3	0	12	<b>26</b>	0	0	0	0	1	0	0	39
	a4	0	0	3	<b>23</b>	0	6	1	0	0	0	33
	a5	0	0	0	5	<b>30</b>	0	0	1	0	0	36
	a6	4	0	0	2	0	<b>24</b>	2	0	0	0	32
	a7	0	0	0	0	0	0	<b>14</b>	1	0	0	15
	a8	0	7	0	0	0	0	6	<b>27</b>	0	0	40
	a9	6	1	0	0	0	0	5	0	<b>29</b>	0	41
	a10	0	0	0	0	0	0	0	0	0	<b>30</b>	30

## 5 Conclusion

This paper presented a compact online action recognition methodology, based on a novel sequence encoding scheme and a lightweight convolutional neural net-



**Fig. 5.** Evolution of class confidences for action #4, performed by subject #6. After a certain point in time, the correct class (4, Right hand draw X) emerges as the most prevalent class.

work. It is suitable for deployment on hardware with limited capabilities, which can be found even in household robotic systems. Experimentation and cross validation showed this pipeline can be used with state-of-the-art human pose estimation techniques to form efficient end-to-end activity recognition techniques.

Discrepancies in the testing data pointed out that there may be room for improvement, as more novel convolutional neural network architectures emerge. Furthermore, there may be hints of overfitting, as stated in the previous section, which may need to be dealt with.

In our future work, we will include attention models [21][28] in the CNN architecture to further improve the performance of the presented approach. Another direction could also be to focus on decrypting the inherent relationships and similarities between actions as encoded by *SRFs*, in order to evaluate the proposed scheme's performance as an action analysis and assessment tool.

## References

1. Bevacqua, E., Raouzaoui, A., Peters, C., Caridakis, G., Karpouzis, K., Pelachaud, C., Mancini, M.: Multimodal sensing, interpretation and copying of movements by a virtual agent. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Weber, M. (eds.) Perception and Interactive Technologies. pp. 164–174. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
2. Beylkin, G.: Discrete radon transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **35**(2), 162–172 (February 1987). <https://doi.org/10.1109/TASSP.1987.1165108>
3. Canal, G., Escalera, S., Angulo, C.: A real-time human-robot interaction system based on gestures for assistive scenarios. *Computer Vision and Image Understand-*

- ing **149**, 65 – 77 (2016). <https://doi.org/10.1016/j.cviu.2016.03.004>, special issue on Assistive Computer Vision and Robotics - Assistive Solutions for Mobility, Communication and HMI
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR* **abs/1812.08008** (2018), <http://arxiv.org/abs/1812.08008>
  5. Chen, C., Jafari, R., Kehtarnavaz, N.: Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International Conference on Image Processing (ICIP). pp. 168–172 (Sep 2015). <https://doi.org/10.1109/ICIP.2015.7350781>
  6. Chen, M., Nikolaidis, S., Soh, H., Hsu, D., Srinivasa, S.: Planning with trust for human-robot collaboration. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction. p. 307–315. HRI '18, Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3171221.3171264>
  7. Chen, Y., Wu, Q., He, X.: Human action recognition by radon transform. In: 2008 IEEE International Conference on Data Mining Workshops. pp. 862–868 (Dec 2008). <https://doi.org/10.1109/ICDMW.2008.26>
  8. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 1243–1252. ICML'17, JMLR.org (2017)
  9. Goudelis, G., Tsatiris, G., Karpouzis, K., Kollias, S.: 3d cylindrical trace transform based feature extraction for effective human action classification. In: 2017 IEEE Conference on Computational Intelligence and Games (CIG). pp. 96–103 (Aug 2017). <https://doi.org/10.1109/CIG.2017.8080421>
  10. Goudelis, G., Karpouzis, K., Kollias, S.: Exploring trace transform for robust human action recognition. *Pattern Recognition* **46**(12), 3238 – 3248 (2013). <https://doi.org/https://doi.org/10.1016/j.patcog.2013.06.006>
  11. Hamdi, D.E., Elouedi, I., Fathallah, A., Nguyuen, M.K., Hamouda, A.: Combining fingerprints and their radon transform as input to deep learning for a fingerprint classification task. In: 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV). pp. 1448–1453 (Nov 2018). <https://doi.org/10.1109/ICARCV.2018.8581072>
  12. Ji, Y., Yang, Y., Shen, F., Shen, H.T., Li, X.: A survey of human action analysis in hri applications. *IEEE Transactions on Circuits and Systems for Video Technology* pp. 1–1 (2019). <https://doi.org/10.1109/TCSVT.2019.2912988>
  13. Kadyrov, A., Petrou, M.: The trace transform and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(8), 811–828 (Aug 2001). <https://doi.org/10.1109/34.946986>
  14. Laptev, I.: On space-time interest points. *International journal of computer vision* **64**(2-3), 107–123 (2005)
  15. Liu, J., Wang, G., Duan, L., Abdiyeva, K., Kot, A.C.: Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing* **27**(4), 1586–1599 (April 2018). <https://doi.org/10.1109/TIP.2017.2785279>
  16. Liu, Y., Nie, L., Liu, L., Rosenblum, D.S.: From action to activity: Sensor-based activity recognition. *Neurocomputing* **181**, 108 – 115 (2016). <https://doi.org/10.1016/j.neucom.2015.08.096>, big Data Driven Intelligent Transportation Systems
  17. Loncaric, S.: A survey of shape analysis techniques. *Pattern Recognition* **31**(8), 983 – 1001 (1998). [https://doi.org/10.1016/S0031-2023\(97\)00122-2](https://doi.org/10.1016/S0031-2023(97)00122-2)

18. Malatesta, L., Murray, J., Raouzaïou, A., Hiolle, A., Cañamero, L., Karpouzis, K.: Emotion modelling and facial affect recognition in human-computer and human-robot interaction. In: M., D.M.I.C. (ed.) *State of the Art in Face Recognition*, chap. 12. IntechOpen, Rijeka (2009). <https://doi.org/10.5772/6648>, <https://doi.org/10.5772/6648>
19. Nibali, A., He, Z., Morgan, S., Prendergast, L.: 3d human pose estimation with 2d marginal heatmaps. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. pp. 1477–1485 (Jan 2019). <https://doi.org/10.1109/WACV.2019.00162>
20. Peters, C., Castellano, G., Rehm, M., André, E., Raouzaïou, A., Rapantzikos, K., Karpouzis, K., Volpe, G., Camurri, A., Vasalou, A.: *Fundamentals of Agent Perception and Attention Modelling*, pp. 293–319. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
21. Rapantzikos, K., Tsapatsoulis, N., Avrithis, Y., Kollias, S.: Bottom-up spatiotemporal visual attention model for video analysis. *IET Image Processing* **1**, 237–248(11) (June 2007)
22. Rodomagoulakis, I., Kardaris, N., Pitsikalis, V., Mavroudi, E., Katsamanis, A., Tsiami, A., Maragos, P.: Multimodal human action recognition in assistive human-robot interaction. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 2702–2706 (March 2016). <https://doi.org/10.1109/ICASSP.2016.7472168>
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)*, <http://arxiv.org/abs/1409.1556>
24. Tsatiris, G., Karpouzis, K., Kollias, S.: Variance-based shape descriptors for determining the level of expertise of tennis players. In: *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*. pp. 169–172 (Sep 2017). <https://doi.org/10.1109/VS-GAMES.2017.8056591>
25. Tsatiris, G., Karpouzis, K., Kollias, S.: Applications of human action analysis and recognition on wireless network infrastructures: State of the art and real world challenges. In: *2018 Innovations in Intelligent Systems and Applications (INISTA)*. pp. 1–6 (July 2018). <https://doi.org/10.1109/INISTA.2018.8466317>
26. Uddin, M.Z., Khaksar, W., Torresen, J.: Activity recognition using deep recurrent neural network on translation and scale-invariant features. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. pp. 475–479 (Oct 2018). <https://doi.org/10.1109/ICIP.2018.8451319>
27. Varia, C., Tsatiris, G., Karpouzis, K., Kollias, S.: A refined 3d dataset for the analysis of player actions in exertion games. In: *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. pp. 1–4 (Aug 2018). <https://doi.org/10.1109/CIG.2018.8490458>
28. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)*
29. Yu, M., Gong, L., Kollias, S.: Computer vision based fall detection by a convolutional neural network. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. p. 416–420. *ICMI '17*, Association for Computing Machinery, New York, NY, USA (2017). <https://doi.org/10.1145/3136755.3136802>