AFRL-AFOSR-UK-TR-2024-0015

Cognitively-inspired architectures for human motion understanding

Nicoletta Noceti
UNIVERSIT DEGLI STUDI DI GENOVA
VIA BALBI 5
GENOVA, , 16126
IT

01/09/2024
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE 20240109	**2. REPORT TYPE** Final	**3. DATES COVERED** **START DATE** 20200930	**END DATE** 20230929

4. TITLE AND SUBTITLE
Cognitively-inspired architectures for human motion understanding

5a. CONTRACT NUMBER	**5b. GRANT NUMBER** FA8655-20-1-7035	**5c. PROGRAM ELEMENT NUMBER**
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Nicoletta Noceti

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSIT DEGLI STUDI DI GENOVA
VIA BALBI 5
GENOVA 16126
IT

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
EOARD
UNIT 4515
APO AE 09421-4515

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-UK-TR-2024-0015

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In the last decades, modelling and understanding human motion from videos has gained an increasing importance in several application domains, including Human-Machine Interaction, gaming, assisted living and robotics. Although the significant advances of the last years, where as in other domains deep learning techniques has gained momentum, the tasks remain among the most challenging, for the intrinsic complexity of dynamic information, and still a lot of work needs to be done to approaching human performance. The biological perceptual systems remain the gold standard for efficient, flexible, and accurate performance across a wide range of complex real-world tasks. A natural inspiration for computational models are thus the mechanisms underlying human motion perception, and the knowledge derived from the Cognitive and Neuro Science fields. Previous works demonstrated the effectiveness of biologically-inspired visual features for object or action recognition , while examples of cognitively-inspired architectures are less present.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT** SAR	**18. NUMBER OF PAGES** 7
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U		

19a. NAME OF RESPONSIBLE PERSON NANDINI IYER	**19b. PHONE NUMBER** *(Include area code)* 314-235-6161

Cognitively-inspired architectures for human motion understanding
FA8655-20-1-7035
Final Technical Report

Principal investigator: Nicoletta Noceti
Host institution: Università di Genova, Italy
Program Officer: Nandini Iyer, European Office of Aerospace Research & Development (AFRL/AFOSR/IOE)

1 Accomplishments

The goal of this project was to investigate the understanding of visual motion information, using models inspired by the human development of such skills.
With a high interdisciplinarity, the project lies at the intersection of Computer Vision, Machine Learning, and Cognitive Science, and is structured according to three layers of analysis
Low-level analysis: *estimating motion saliency from image sequences*. The goal is the identification of relevant parts on which further processing can be focused. Particular attention is on methods able to offer a compromise between effectiveness and efficiency. This goal has been mainly developed during the first and second year of activity.
Mid-level analysis: *representing motion using primitives*. The goal is to study strategies for detecting and combining human motion primitives to be organised in hierarchical descriptions for actions and activities representation. This activity spanned the three years of the project.
High-level analysis: *learning motion models*. The goal is to study approaches for transferring knowledge from one task to another, controlling the amount of data and the level of supervision that is needed, and the use of contextual information for action recognition and, possibly, anticipation. As the previous one, this activity spanned the three years of the project.

During the third and last year of this project, the ongoing activities have been finalized or strengthened. Technical details on such activities will be reported in Sec. 4, while here we summarize the main research directions. We devoted a large part of our activity to the design and development of motion representation and understanding approaches leveraging the notion of primitive and the metaphor of an action being a sentence composed of motion words. To this purpose, we explored the use of Transformers, an approach for representing sequential data largely used in these years. In addition, we considered the use of scene graphs to model the connections between humans and the environment to recognize and anticipate the prediction of more structured activities. Finally, anticipation has been also addressed in collaboration with the IMAGINE team at École des Ponts ParisTech. Being the main research directions of the last period in the project, the papers on this part are currently in preparation.
A second main direction of activity (in strong collaboration with AMAZON Aws, DKFZ, ISTA and the University of Chicago) considered the causality paradigm, and in particular the design and analysis of efficient and robust causal discovery approaches and disentangled representations ([21, 22, 19], one paper under review [20], one in preparation). Finally, we finalized the work which has been mainly developed during the first two years of the project. On motion saliency estimation, we analysed the robustness of our approaches to scene bias [27]. On the estimation of head direction and its use in the context of social interaction analysis, we are exploring its use in the robotics domain in collaboration with IIT (a paper in preparation).

2 Impacts

The full list of publications in the three years of the project is [24, 25, 26, 27, 23, 9, 6, 7, 8, 5, 2, 16, 21, 22, 19, 20]. In addition, 2 papers are currently under review (on the CVIU journal and the AISTAT conference) and 5 papers are in preparation.

From a theoretical point of view, our research has an impact in the field of human motion analysis from visual data, using methods with controlled complexity and interpretability, able at the same time to provide significant results. On the application side, a natural application domain has been social interaction analysis and more in general Human-Machine/Robot interaction, where the ability to have artificial agents able to perceive (and possibly react, although this is out of the scope of the project) the visual stimuli is paramount for a pleasant and successful interaction. The experimental validation of many of the methods developed in the project in the robotic field has been possible thanks to the collaboration with the Istituto Italiano di Tecnologia.

Inspirations from the present project and its achievements have been capitalised in two newly funded projects. RAISE (*Robotics and AI for socio-economic empowerment*) has been funded by the EU, through the Italian Ministry of University and Research and more specifically the National Recovery and Resilience Plan. This initiative aims to create and strengthen innovation ecosystems in areas of technological specialization consistent with the industrial and research vocations of the reference territory. In the project, we further develop our approaches for human motion analysis and recognition with particular emphasis on rehabilitation, assistive living and social interaction analysis.

Moreover, a new project has been funded by AFOSR, titled *Multi-modal learning for gait-based human analysis and authentication*. The goal of the project is to strengthen the ability of biometric authentication through gait using portable devices like smartphones in challenging conditions of data acquisition. The key idea is to exploit transfer and multi-modal learning across different data modalities, in particular from visual data to time sequences acquired with accelerometers and gyroscopes.

3 Changes

The developed activity follows the plan originally included in the proposal, with two main additions – both of them already mentioned in the technical reports of the past two years. The first one regards the use of Head Orientation as a cue for human attention and social interaction, and ultimately a way to help motion recognition by adding contextual human-centered information.

The second addition is the one related to the study of the causality paradigm, which has been a secondary activity but shows a high potential. Although not mentioned in the original proposal, the activity started during the second year of the project thanks to a new, fruitful collaboration with leading researchers in the field. The study and exploitation of the causality paradigm in the field of human motion analysis from visual data is still largely unexplored, while the presence of time that naturally leads to cause-effect relationships in the observed data suggests it may be a very promising application domain. This is the main reason behind the choice of including this line of research in the project.

4 Tecnical Updates

In the third and last year of the project, we mainly focused on human motion understanding. A second line of activity was investigating the causality paradigm. In this section, we sketch the main activities.

Minimal action representation: using motion primitives

Nowadays, motion-based tasks such as action or activity classification from videos are addressed using end-to-end deep architectures, able to establish a connection between video data and the high-level concept to be learned [1]. Although the astonishing performance they can achieve, some drawbacks must be mentioned. First, given the complexity of the model, such methods are highly data-hungry, with a consequent severe load on the training phase. Second, they are sort of black box, where interpretability and explainability of the derived models are not facilitated [14].

We follow a different inspiration, where the tasks of motion representation and understanding are partially decoupled. A building block of biological motion understanding is the identification of motion primitives [33]. During the last year, we strengthened our work on hierarchical motion descriptions that leverage the notion of motion primitive. In our approach, a primitive corresponds to a portion of the motion stream characterized by the presence of a velocity bell, a known property of human motion. Indeed, it has been shown that humans, if explicitly asked, tend to segment motion in correspondence with points with low kinematics [11]. For deriving the velocity, we start from the output of pose detectors as [18, 3, 29] and approximate the velocity as differences between adjacent positions in time. We encode different concepts of motion – a gesture, an action, an activity – as a sequence of primitives and in a hierarchical way (a gesture is encoded as a sequence of primitives, an action as a sequence of gestures, an activity as a sequence of actions) so to favour the sharing of information

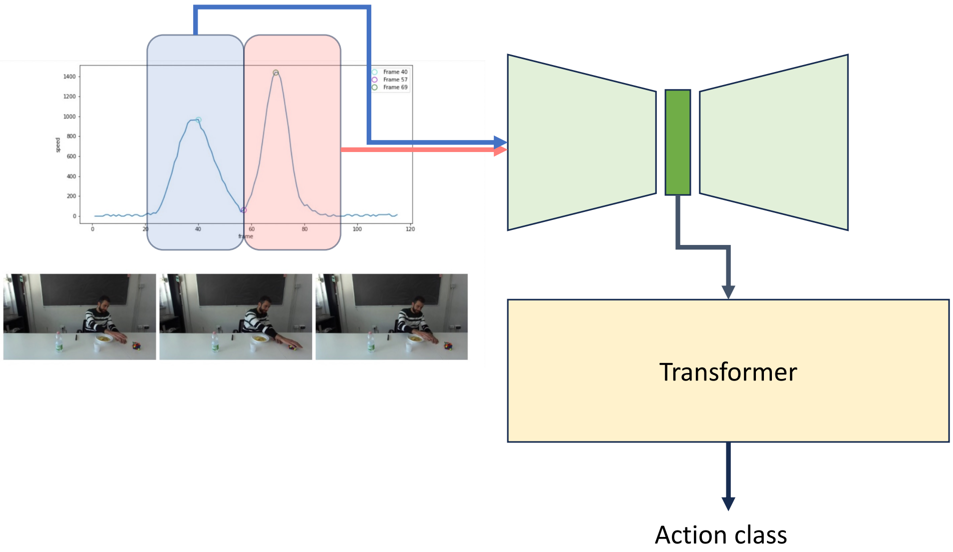

Figure 1: A sketch of our pipeline for human motion understanding. We start by segmenting the joints motion over time in primitives, that are projected into an embedding using a Variational Autoencoder. The sequence of *motion words* derived in this way represents the *motion sentence* we provide in input to a Transformer to get the final motion classification. This hierarchical approach is suited for classifying different concepts of motion, from gestures, to actions, to more structured activities.

between different tasks.

As sketched in Fig.1 we treat an action (the sentence) as a sequence of primitives (words) and we exploit the use of self-attention [15, 34] to learn a representation able to put a primitive in the context of the action. An important pre-processing step is based on an encoding of motion primitives for each body joint, for which we adopt a Variational Autoencoder.

Experiments are currently ongoing also on a big dataset [30] for two publications and the PhD thesis of F. Figari Tomenotti.

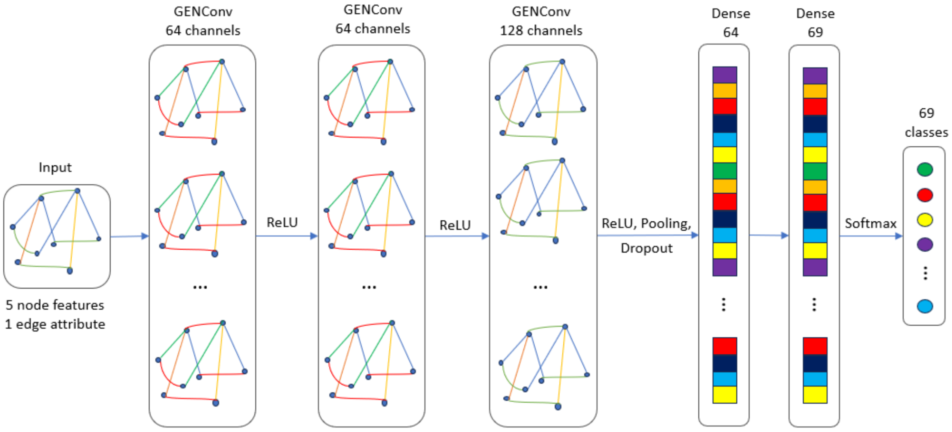

Figure 2: Our pipeline for motion classification based on scene graphs. Each frame is compactly represented by an undirected graph (i.e. the scene graph, see Fig. 3). Such graphs, stacked over time, compose the spatio-temporal graph that represents the input to our network. The latter has been designed from scratch exploiting graph convolutions, targeting a high representation power while keeping the complexity under control. A sequence of dense layers forms the final classifier.

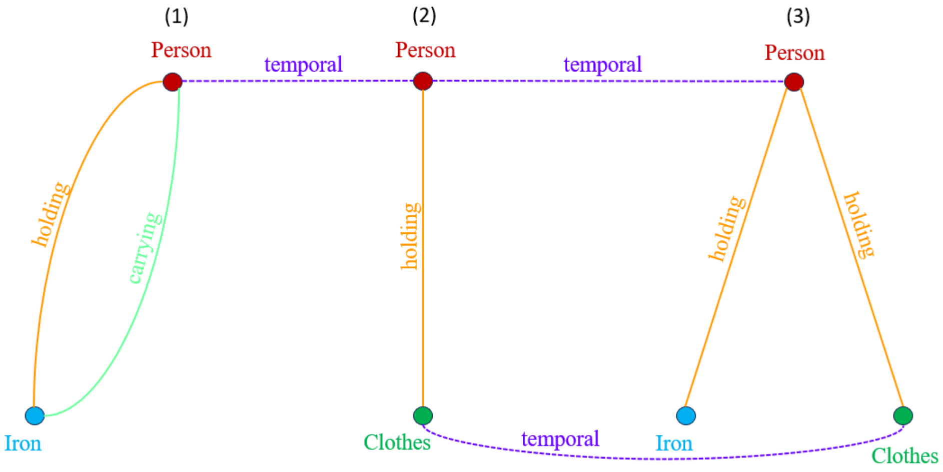

Figure 3: In a scene graph each element of the scene (including humans) is a node, while relations between elements are encoded with edges. Both nodes and edges are labelled with relevant information depending on the task. The spatio-temporal scene graph is obtained by adding temporal edges to explicitly connect the same elements in different time instants.

Scene graphs and activity recognition

Anticipation of the actions and the intentions of other people is a main cue we exploit during human-human social interaction. Given its importance, it is of great interest to devise computational methods able to provide similar abilities to artificial agents.
During the first years of the project, we considered Head Pose as a visual cue that can provide relevant indications of the people's intentions. In this last year, we considered the role of the context and the environment.
In Fig. 2 we report a visual representation of our method. We approached the problem by encoding the action context with a scene graph [13, 12, 28], that captures the connection between the human and the structural elements of the environment. Humans and scene elements are nodes of an undirected graph, while the edges represent the relations between them, which can be of various types (e.g. geometrical, manipulation, attention,...). The graphs in sequence provide a spatio-temporal representation of the actions/activity and can be provided in input to a deep architecture appropriately designed. In Fig. 3 we show an example of an evolving activity where a person is manipulating an iron and clothes. As for the architecture, we opted for a Graph Convolutional Neural Network designed from scratch for the purpose, in order to balance the complexity and the representation capability. We experimentally analysed the method on both recognition and anticipation tasks on the Home Action genome dataset [31]. The dataset includes 69 daily activities involving 86 object classes (including the class Person) and 25 types of relationships.
The results are very satisfactory: we showed that the method can reach 79% of accuracy when the whole action/activity is observed, but it is also very robust to a reduction in the observation time (for anticipation): when observing about the 50% of the action the accuracy is still above the 70%. We also noticed that for both the recognition and anticipation tasks the main sources of errors are due to actions involving the same atomic activities but developed in different orders. We argue this may be solved by using directed graphs, which will be an objective of our future investigation.
A paper is currently in preparation for this activity.

Actions anticipation

Prediction abilities of methods have been also the topic of interest within a collaboration with the IMAGINE team of the École des Ponts ParisTech. More specifically, we worked on predicting actions in the following setting. Given the observation of a (possibly short) amount of video depicting an activity (i.e. a sequence of actions), the task is to predict a certain amount of future actions conditioned by the activity. In the observation part, the model should be able to classify the actions in progress, before predicting the labels for the future. The task is described in a challenge proposed in the CVPR conference in 2022 coupled with the Ego4d dataset [10], and it has only ego-vision videos acquired in many different settings by many people from around the world. This setting represents an interesting application scenario for our approaches, since in videos acquired

from ego vision often the human body is poorly visible, providing very challenging working conditions.
As shown in Fig. 4 we set up an experimental pipeline based on two Transformer architectures: with the first one, features are derived from the video [4], while the second (designed from scratch) translates such features to action features. The latter can be used for both the action(s) recognition and the action(s) anticipation task. The work is currently on hold, but the experiments will be finalized in the next months for a paper submission.

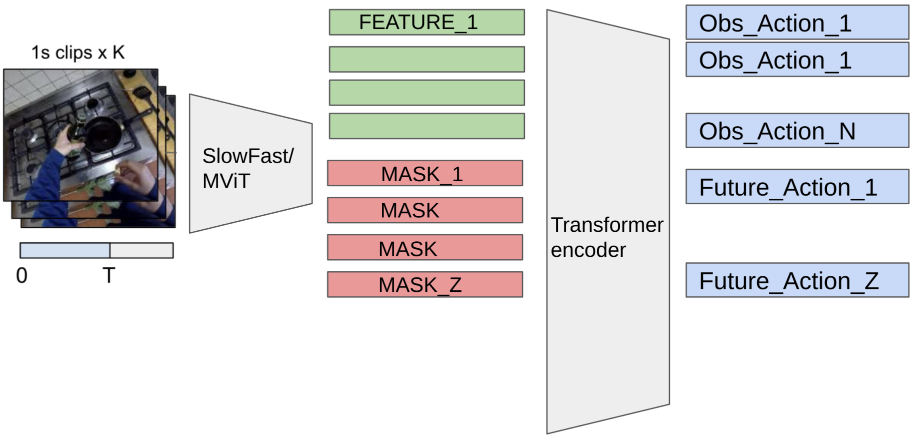

Figure 4: A visual representation of our pipeline for recognizing and predicting (i.e. anticipating) sequences of actions forming a more structured activity. A Multiscale Vision Transformer is used to learn relevant features from the video. For this purpose, the video is segmented into clips. Then, the sequence of representations of such clips is injected into a second Transformer (which we designed from scratch) to get action(s) encoding and the final labels. The latter include a sequence of labels of the observed actions (recognition) and a sequence of labels of the expected future actions (prediction, or anticipation).

Causal discovery

In the last few years, causality has emerged as a powerful and elegant framework for machine learning models. Of particular interest for this project is the ability to improve generalisation, not only interpreted in the classical way (from one data point to the next) but also in a harder sense, from one task to another one [32]. Indeed, in a modular representation of the world (where the modules correspond to physical causal mechanisms), many modules are expected to behave similarly across different tasks and environments. When learning a causal model, fewer examples may be required to adapt to a new task/environment, as most knowledge (the modules) can be reused without further training
In collaboration with experts in the field of causality (including AMAZON AWS in Tuebingen, ISTA and the University of Chicago) we worked on causal discovery, with the specific aim of devising methods that provide effectiveness, robustness to challenging conditions and computational efficiency. In particular, in [19] we extensively benchmark the empirical performance of recent causal discovery methods (including the approach we proposed in [22, 21]) on observational iid data generated under different background conditions, allowing for violations of the critical assumptions required by each selected approach. These evaluations are paramount for the application of such methodologies on real-world tasks. Furthermore, we discussed in [20] the criticalities in the evaluation of causal discovery methods on synthetic data.
In a parallel activity, we studied and evaluated disentangled representations on visual data. An intuition on disentanglement is that it refers to the separation in different parts of the learned representations of the main factors of variation that are present in the data distribution [17]. From a theoretical point of view, they may provide a boost to the models' ability to generalize to unseen domains. However, from a more practical viewpoint, the task is still largely unexplored, and its applicability in real scenarios is still debated.
As a follow-up of this activity, we are currently investigating the use of causality for sequential data, to exploit such mechanisms in the tasks of interest for a new project funded by AFOSR, titled *Multi-modal learning for gait-based human analysis and authentication*.

References

[1] M. Asadi-Aghbolaghi and et al. Deep learning for action and gesture recognition in image sequences: A survey. In *Gesture Recog.*, pages 539–578. Springer, 2017.

[2] G. Cantarini, F. F. Tomenotti, N. Noceti, and F. Odone. Hhp-net: A light heteroscedastic neural network for head pose estimation with uncertainty. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3521–3530, 2022.

[3] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[4] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

[5] L. Garello, L. Lastrico, F. Rea, F. Mastrogiovanni, N. Noceti, and A. Sciutti. Property-aware robot object manipulation: a generative approach. In *2021 IEEE International Conference on Development and Learning (ICDL)*, pages 1–7. IEEE, 2021.

[6] L. Garello, F. Rea, N. Noceti, and A. Sciutti. Towards third-person visual imitation learning using generative adversarial networks. In *2022 IEEE International Conference on Development and Learning (ICDL)*, pages 121–126. IEEE, 2022.

[7] L. Garello, F. Rea, A. Sciutti, and N. Noceti. Embodied generative third person view translation and encoding. *I-RIM*, 2020.

[8] L. Garello, F. Rea, A. Sciutti, and N. Noceti. A generative model towards conditioned robotic object manipulation. *I-RIM*, 2021.

[9] G. Goyal, N. Noceti, and F. Odone. Cross-view action recognition with small-scale datasets. *Image and Vision Computing*, 120:104403, 2022.

[10] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[11] P. E. Hemeren and S. Thill. Deriving motor primitives through action segmentation. *Frontiers in psychology*, 1:243, 2011.

[12] J. Johnson, A. Gupta, and L. Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[13] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[14] O. Li, H. Liu, C. Chen, and C. Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[15] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[16] L.Lastrico, L.Garello, F.Rea, N.Noceti, F.Mastrogiovanni, A.Sciutti, and A.Carfí. Robots with different embodiments can express and influence carefulness in object manipulation. In *Proceedings of the International Conference on Developmental Learning and Epigenetic Robotics (to appear)*, 2022.

[17] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[18] G. H. Martinez, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, and Y. Sheikh. Single-network whole-body pose estimation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6981–6990. IEEE, 2019.

[19] F. Montagna, A. A. Mastakouri, E. Eulig, N. Noceti, L. Rosasco, D. Janzing, B. Aragam, and F. Locatello. Assumption violations in causal discovery and the robustness of score matching. In *Neurips*, 2023.

[20] F. Montagna, N. Noceti, L. Rosasco, and F. Locatello. Shortcuts for causal discovery of nonlinear models by score matching. *arXiv preprint arXiv:2310.14246*, 2023.

[21] F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *CLEAR*, 2023.

[22] F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Scalable causal discovery with score matching. In *CLEAR*, 2023.

[23] V. Nair, P. Hemeren, A. Vignolo, N. Noceti, E. Nicora, A. Sciutti, F. Rea, E. Billing, F. Odone, and G. Sandini. Action similarity judgment based on kinematic primitives. In *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 1–8. IEEE, 2020.

[24] E. Nicora and N. Noceti. Salient motion detection and representation using efficient projection kernels. In *PERCEPTION*, volume 50, pages 54–54. SAGE PUBLICATIONS LTD 1 OLIVERS YARD, 55 CITY ROAD, LONDON EC1Y 1SP, ENGLAND, 2021.

[25] E. Nicora and N. Noceti. Exploring the use of efficient projection kernels for motion saliency estimation. In *International Conference on Image Analysis and Processing*, pages 158–169. Springer, 2022.

[26] E. Nicora and N. Noceti. On the use of efficient projection kernels for motion-based visual saliency estimation. *Frontiers in Computer Science*, 4:867289, 2022.

[27] E. Nicora, V. P. Pastore, and N. Noceti. Gck-maps: A scene unbiased representation for efficient human action recognition. In *International Conference on Image Analysis and Processing*, pages 62–73. Springer, 2023.

[28] J. Ost, F. Mannan, N. Thuerey, J. Knodt, and F. Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021.

[29] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018.

[30] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021.

[31] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J. C. Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2021.

[32] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[33] F. Stulp, E. A. Theodorou, and S. Schaal. Reinforcement learning with sequences of motion primitives for robust manipulation. *IEEE Transactions on robotics*, 28(6):1360–1370, 2012.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.