Deep Learning for Assistive Computer Vision

Marco Leo[1(✉)], Antonino Furnari[2], Gerard G. Medioni[3], Mohan Trivedi[4], and Giovanni M. Farinella[2]

[1] Institute of Applied Sciences and Intelligent Systems, National Research Council of Italy, Lecce, Italy
marco.leo@cnr.it
[2] University of Catania, Catania, Italy
{furnari,gfarinella}@dmi.unict.it
[3] University of Southern California, Los Angeles, USA
medioni@usc.edu
[4] University of California, Oakland, USA
mtrivedi@soe.ucsd.edu

Abstract. This paper revises the main advances in assistive computer vision recently fostered by deep learning. To this aim, we first discuss how the application of deep learning in computer vision has contributed to the development of assistive techinologies, then analyze the recent advances in assistive technologies achieved in five main areas, namely, object classification and localization, scene understanding, human pose estimation and tracking, action/event recognition and anticipation. The paper is concluded with a discussion and insights for future directions.

Keywords: Assistive technologies · Computer vision · Deep learning

1 Introduction

Computer vision is attracting more and more people coming from academia and industry. In the context of Assistive Technologies (AT), researches have already proved how computer vision algorithms can be effectively exploited to address different user's needs pointed by the World Health Organization (e.g., Mental Function, Mobility, Sensory Substitution and Assisted Living) [20]. It is straightforward to figure out that the opportunities to address more challenging assistive tasks depend on the speed with which the fundamentals of knowledge evolve, as commonly happens when knowledge is transferred from theoretical to application fields.

After a decade in which visual intelligence performances remained quite stable, over the past 5 years or so, it has been pushed up by the impact of the application of the deep learning paradigm. This has gone from a somewhat niche field comprised of a strict group of researchers to being mainstream and enabling breakthrough applications of diverse areas such as image and video understanding, speech recognition, medical imaging and self driving vehicles. The idea of

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11134, pp. 3–14, 2019.
https://doi.org/10.1007/978-3-030-11024-6_1

deep learning dates back to the end of 80's when Neural Networks (NN) began to be used for mapping inputs to outputs with the aim to automatically recognize handwritten characters.

In the mid of the 90's, the interest of the computer vision community in NN decreased since the architectures did not allow to be scaled to complex contexts where other methods, such as Kernel Machines based on linear classifiers (e.g., SVM), were more effective in most of the application domains involving visual recognition tasks. This scenario changed when large visual datasets and powerful computational resources became available at hand, making possible the increase of the complexity of architecture based on neural networks, as well as the learning capabilities, thus moving the focus from shallow machine learning to deep machine learning. Indeed, supervised deep learning methods thrive on big datasets, which before the ImageNet era were only available for some specific tasks (e.g., handwriting recognition). Nowadays, datasets for image understanding have become big enough to train deep learning systems, and modern GPUs allow researchers to implement effective algorithms which beat almost any record in computer vision. This fact has gained the attention not only of the computer vision community, but also the one of other scientific research fields.

One of the greatest advantage of deep learning is the possibility to learn effective representations of the data for a given task [14]. For this reason, deep learning is currently considered the primary candidate for any visual recognition task [38] and it is being extended to visual reasoning [17]. This is made evident by the availability of several books which introduce and discuss the different deep learning approaches, as well as by the number of papers describing its use in different application fields such us medical imaging [22], health informatics [31], feature learning [50], etc. Assistive technologies do not escape this trend since they have been already flooded by deep learning. An up-to-date overview on different types of deep neural networks and recent progresses is given in [23] where also applications of deep learning techniques on some selected areas (speech recognition, pattern recognition and computer vision) are highlighted. This paper tries to summarize instead how deep learning has been recently exploited to deal with assistive tasks. The rest of the paper is organized as follows. Section 2 introduces a taxonomy concerning the way in which deep learning influenced assistive technologies. Section 3 discusses recent works in the context of assistive computer vision. Finally, Sect. 4 gives some hints for possible future directions.

2 How Deep Learning in CV Is Being Strengthening and Improving AT

The task of automatically recognizing and locating objects is one of the primary tasks for humans in order to survive, work and communicate. As a consequence, the ability to automatically perform this task starting from images and videos is fundamental to build very powerful assistive devices able to understand and/or interact with their surroundings and, as a consequence, to help people with cognitive and/or physical limitations. On the other hand, deep architectures can

learn more complex models than shallow ones, since they learn powerful representations of the objects without the need to perform hand design features. The deep learning frameworks for object recognition methods can mainly be categorized into two groups: one follows the traditional object detection pipeline, involving the generation of region proposals and the classification of each proposal into different object categories. The other regards object detection as a regression or classification problem, adopting a unified framework to achieve final results (categories and locations) directly. Recent advances in this area can be found in [52].

A strictly related task is the one so called scene understanding, i.e. the ability not only to identify the targets (as object recognition does), but also to understand the other properties of the observed scene. In other words, this task entails recognizing the semantic constituents of a scene and the complex interactions that occur between them. Humans have no difficulty with these tasks and can associate semantic information with the scene at different levels. This challenging task can be approached in different ways. At a very high level, the approaches can be divided into two main categories: using low-level features, and using object recognition. However, many other techniques are integrated into each of these approaches, including probabilistic, and/or fuzzy techniques, in order to deal with the uncertainty which often attends the result of image understanding. Convolutional Neural Networks (CNNs) can be really useful also to solve this task. A recent approach to address the aforementioned challenge consists in using the convolutional patch networks, which are CNNs trained to distinguish different image patches giving the possibility to perform pixel-wise labeling [5]. One of the bottlenecks in training for better representations is the amount of available per-pixel ground truth data that is required for core scene understanding tasks such as semantic segmentation, normal prediction, and object boundary detection. To address this problem, a number of works proposed using synthetic data and some of them also provide a systematic study of how such synthetic data is generated [51].

Many assistive technologies aim at assisting people in overcoming physical and cognitive barriers by tracking their body pose or by recognizing their activity and their actions. Different approaches using deep learning have been recently proposed and, among all, the one based on ensemble of models, each of which is optimized for a limited variety of poses, is capable of modeling a large variety of human body configurations [19]. Alternative approaches rely on inferring the dependencies between human joints that are modelled via a max-margin structured learning framework [18]. However tracking multiple people in realistic videos is still an open research area in which the introduction of new large-scale benchmarks is helping to build increasingly performing models [3]. A related research topic concerns the recognition of an event that is a conceptually higher semantic problem that could capture the complex behavior of a group of people, interacting with multiple objects, and taking place in a specific environment. Transferring deep object and scene representations for event recognition is a very smart solution recently proposed in [46]. The recognition and

understanding of these tasks are fundamental in human-robot interactions, where there is also a need for the machine to make decisions based on the understanding of the near future, i.e to anticipate the next event [7] or object [12]. This ability comes naturally to us and we make use of it subconsciously. Almost all human interactions rely on this action-anticipation capability. The ability to anticipate the action of other individuals is essential for our social life and even survival. Therefore, it is critical to transfer these abilities to computers, and these challenging tasks have been strongly pushed forward in knowledge thanks to Deep Learning. By using CNNs it is possible to predict human actions only observing a few frames of a video containing an action [34]. Exploiting an encoder-decoder recurrent neural network to address the action prediction problem, it is possible to predict multiple and variable-length action sequences [36]. Of course there are many other areas of computer vision that are receiving a strong impact from the development of the deep learning paradigm. However, the tasks just mentioned are the foundations for many assistive frameworks, as we will see in the next section, and therefore in this paper we will limit the discussion to them.

3 Recent Advances in at Exploiting DL Strategies

In the following subsections we will analyze how the recent deep learning based advances in computer vision tasks, described in Sect. 2, have been exploited to improve AT frameworks. As already anticipated, the following tasks will be considered:

- Object localization and recognition;
- Scene understanding;
- Human pose estimation and tracking;
- Action and event recognition;
- Anticipation.

3.1 Object Localization and Recognition

Object Localization and Recognition is one of the areas of computer vision that is maturing very rapidly thanks to deep learning. Nowadays, there is a plethora of pre-trained deep learning models which can be used for this task, so it only takes a small amount of effort to build a system able to detect most of the objects in an image or video even in the presence of multiple overlapping objects and different backgrounds. In addition to detecting even multiple objects in a scene, recent deep learning based architectures are also able to precisely identify their boundaries and relations to one another. This is achieved by deep structured learning which, for example, can learn relationship by using both feature, geometry, label [53] and, even physics and inferences about the abstract properties of the whole system [4].

The recent advantages in object localization and recognition have been already employed in different AT applications. For instance CNNs are effectively exploited to improve the performance in the autonomous navigation [8].

In [25] the YOLOv2 engine [32], which is one of the fastest strategies for object detection, is used to improve interaction with other subjects within an indoor navigation system that guides a user from point A to point B with high accuracy. The Yolov2 engine was recently used also in [16] to build a multimodal computer vision framework for human assistive robotics with the purpose of giving accessibility to persons with disabilities. Recently, object detection and recognition has been exploited to build a marker-less model for motion-aware gait assessment by localizing feet in egocentric videos [26].

3.2 Scene Understanding

At low level, this concerns the detection of structures of the scene, such as the task of finding edges arising from the physical surfaces of a scene. The extraction of useful scene information can underpin many computer vision tasks such as sketch recognition and 3D scene reconstruction, and is important for conveying knowledge for assistive navigation systems. This is also a key task in retinal implants. Improving the recovery of structural edges using RGB-D input was addressed in [10] with an end-to-end fully convolutional neural network approach. At the higher level, scene understanding concerns the ability of vision systems to infer and describe the content of the scene. As example, in [48] it is proposed a visual question answering system based on DL designed around spoken questions asked by blind people about their surroundings using a mobile phone camera picture. Another example of high level scene understanding regards the ability of a vision system to recognize locations of interest for a user in order to perform temporal segmentation of videos for lifelogging applications [11]. Scene understanding also plays a key role in automatic story comprehension that trough CNN can provide effective solutions for the visually impaired or cognitive robotics [42].

3.3 Human Pose Estimation and Tracking

The estimation of the articulated motion of the human body is useful for a number of real world applications including medical rehabilitation, human-robot interaction and in general to create smart environments suitable to understand people behaviours.

Pose estimation is generally pursued by detecting and extracting the positions of the joints of the human body from different sources such as a single image, a sequence of images, and RGB-D data. The main goal is to reconstruct the skeletal structures of the people in the scene and hence provide information about their body posture, the motion of the body, and human gestures. Understanding human poses from images is considered one of the major challenges in the field of Computer Vision and has been intensively studied in the last few decades by the research community. A propulsion in this research field has been given by the work reported in [43], where the problem of pose estimation is formulated as a regression problem to infer the position of the joints of the body in a Deep Neural Network framework. In the context of assistive technologies,

monitoring the pose of a child over time could reveal important information both during clinical trials [21] or natural behaviors [40]. Human pose estimation methods have been tested on a variety of challenging conditions, but few studies to highlight performance specifically on children's poses have been done. Infants, toddlers and children are not only smaller than adults, but also significantly different in anatomical proportions. In [37] is proposed a study in which different deep learning based approaches for human pose estimation are compared when subjects are children. Results reveal that accuracy of the state of art methods drops significantly, opening new challenges for the research community. The pose of humans can be also used to understand emergency situations, such as unintended falls of an elderly person which lives alone. Fall recognition can be treated as a binary classification problem to obtain frame-wise semantic labels, so that fall recognition and its localization in time can be addressed simultaneously. In the recent literature different methods based on convolutional neural networks using both, RGB and RGB-D data have been proposed to address the fall recognition problem [15].

3.4 Action and Event Recognition

The action recognition task is related to the identification of the different possible actions performed by a human from a sequence of frames, where the actions may or may not be performed throughout the entire duration of the video. Generally speaking an action can be regarded to as a temporal evolution of some visual features. Hence the task becomes to model this evolution to recognize the occurrence of actions.

Common examples of actions to be recognized are "answer phone", "shake hands" (Short actions), "make sandwich", "do homework" (Activities/events with one actor), "birthday party", "parade" (Activities/events involving several persons), but also facial actions such as "smile". Current state-of-the-art approaches for spatio-temporal action localization [45] rely on detections at the frame level and model the temporal context with 3D ConvNets. Advanced approaches model spatio-temporal relations to capture the interactions between human actors, relevant objects and key scene elements to discriminate among human actions [41].

Action recognition in first person vision (proprioceptive activity recognition) is a new frontier of research with a series of additional challenges with respect to classic action recognition.

First-person vision (FPV) activity recognition involves the use of wearable cameras and is greatly beneficial for assisted living, life-logging and summarization [27]. In this context, motion representations which use stacked spectrograms have been proposed in [1]. These spectrograms were generated over temporal windows from mean grid-optical-flow vectors and the displacement vectors of the intensity centroid. The stacked representation enables the system to use 2D convolutions to learn and extract global motion features. Moreover, a long short-term memory (LSTM) network was used to encode the temporal dependency among consecutive samples recursively.

In [30], a deep learning model useful to predict the next action task to be performed by a robot is proposed. The model exploits both, the recognition of objects and their relations. The preconditions and effects of the robot actions are modeled through symbolic language, and the next goal state learning is obtained with a multi-layered LSTM architecture fed by the predicates with terms verified by vision.

One of the main problems with the use of assistive visual monitoring systems in the wild is the requirement of a large amount of training data for each new environment, as models trained in one location tend not to generalize well to others. If improvements could be found by leveraging existing data to circumvent or at least speed the training process in new environments, the deployment of such systems could become faster and easier, enabling more widespread use and providing robust results. In [29], the issue of transfer learning for frame-based event classification using RNNs was tackled.

3.5 Anticipation

The ability to anticipate future events is a desirable capability for assistive technologies. Algorithms to anticipate future events in order to support automated decisions and assist the user have been recently investigated by the computer vision research community. Some efforts have focused on the use of egocentric cameras, which allow to acquire video from the point of view of the user, in order to infer their future actions. Among these works, in [39] it is proposed a method able to anticipate the next action likely to be performed by a user from egocentric video and infer whether that action is correct in the work-flow. The authors of [28] used CNNs to predict the future location of the camera wearer in an egocentric video. The study in [49] designed a method based on Generative Adversarial Networks (GAN) to infer the gaze of the user in future frames (e.g., to infer what the user will observe next). In [9] it is proposed a deep learning architecture to anticipate the position of specific objects and hands in future frames, whereas the study reported in [33] used inverse reinforcement learning to understand the user's goal and anticipate the next location and object they will be interested in. The authors of [7] proposed the task of anticipating future actions performed by the camera wearer on a newly proposed large dataset of videos of egocentric activities in kitchens.

Other works investigated anticipation tasks in the context of third person vision. For instance, in [44] it is proposed to use deep convolutional networks to anticipate multiple future representation from the current frame of a video. The anticipated representations were then used to forecast future actions and objects. In [13] it is proposed an encoder-decoder LSTM architecture capable of anticipating future representation and predict future actions. The study in [24] describes a systems to predict future actions and their starting time. The authors of [2] investigated two deep models to predict multiple future actions and their duration from video.

4 Future Directions

The possible topics most likely to be explored thanks to deep learning will be oriented towards the effective modeling of human behaviors and cognition. Indeed, these aspects are essential to build adaptation and personalization mechanisms for assistive systems. Since eye gaze has been frequently studied in interactive intelligent systems as a cue for inferring user's internal states and to have priors about the user intent, a significant body of works could investigate the relationships between eye movements and cognitive processes to provide an understanding into memory recall, cognitive load, interest, the level of domain knowledge, problem solving, desire to learn, and strategy use in reasoning. A contribution in this direction has already been proposed in [6], where a CNN-based method was trained to estimate user's gaze fixations on the tablet screen (while answering a question) to automatically gather a set of eye movement features useful to discriminate users knowing the correct answer with respect to the others.

Another very promising research line concerns the ability to automatically discover properties and affordances of regions of scenes (e.g., the affordance of objects), which indicate their relevance for a certain functional interaction with the user. Segmenting affordance regions, however, is more difficult than classical semantic image segmentation, where the focus in more on the objects present in the scene. This means that affordance segmentation requires to predict a set of labels per pixel since an object region might contain multiple affordance types. A weakly supervised semantic image segmentation approach based on deep learning was recently proposed in [35], where it is exploited an adaptive approach for binarizing the predictions of a convolutional neural network.

Anticipation methods will evolve to incorporate long term relationship between the observed events to perform better predictions of the future. This research area will allow to build proactive assistive systems and improve human-machine as well as help to anticipate the interactions between a human and the surrounding objects [12].

For some assistive tasks, the application of deep learning strategies has not yet happened, even if it would be highly required to standardize some critical procedures such as support to early diagnosis or assessment of neurodevelopmental disorders. This is manly due to the lack of publicly available datasets. For example in [47] a dataset containing RGB-D data related to real infant movements with varying realistic textures, shapes and backgrounds has been proposed in order to speed-up the medical infant motion analysis based on the training of deep learning approaches.

Acknowledgments. This research has been supported by Piano della Ricerca 2016–2018 linea di Intervento 2 of DMI, University of Catania.

References

1. Abebe, G., Cavallaro, A.: A long short-term memory convolutional neural network for first-person vision activity recognition. In: Proceedings of International Conference on Computer Vision Workshops (ICCVW) (2017)
2. Abu Farha, Y., Richard, A., Gall, J.: When will you do what?-anticipating temporal occurrences of activities. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5343–5352 (2018)
3. Andriluka, M., et al.: PoseTrack: a benchmark for human pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5167–5176 (2018)
4. Battaglia, P., Pascanu, R., Lai, M., Rezende, D.J., et al.: Interaction networks for learning about objects, relations and physics. In: Advances in Neural Information Processing Systems, pp. 4502–4510 (2016)
5. Brust, C.A., Sickert, S., Simon, M., Rodner, E., Denzler, J.: Efficient convolutional patch networks for scene understanding. In: International Conference on Computer Vision Theory and Applications (VISAPP) (2015)
6. Celiktutan, O., Demiris, Y.: Inferring human knowledgeability from eye gaze in m-learning environments. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018 Workshops, LNCS, vol. 11134, pp. 193–209. Springer, Cham (2019)
7. Damen, D., et al.: Scaling egocentric vision: the EPIC-KITCHENS dataset. arXiv preprint arXiv:1804.02748 (2018)
8. Erol, B.A., Majumdar, A., Lwowski, J., Benavidez, P., Rad, P., Jamshidi, M.: Improved deep neural network object tracking system for applications in home robotics. In: Pedrycz, W., Chen, S.-M. (eds.) Computational Intelligence for Pattern Recognition. SCI, vol. 777, pp. 369–395. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-89629-8_14
9. Fan, C., Lee, J., Ryoo, M.S.: Forecasting hand and object locations in future frames. CoRR abs/1705.07328 (2017). http://arxiv.org/abs/1705.07328
10. Feng, D., Barnes, N., You, S.: DSD: depth structural descriptor for edge-based assistive navigation. In: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 1536–1544. IEEE (2017)
11. Furnari, A., Battiato, S., Farinella, G.M.: Personal-location-based temporal segmentation of egocentric videos for lifelogging applications. J. Vis. Commun. Image Represent. **52**, 1–12 (2018)
12. Furnari, A., Battiato, S., Grauman, K., Farinella, G.M.: Next-active-object prediction from egocentric videos. J. Vis. Commun. Image Represent. **49**, 401–411 (2017)
13. Gao, J., Yang, Z., Nevatia, R.: RED: reinforced encoder-decoder networks for action anticipation. In: British Machine Vision Conference (2017)
14. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016). http://www.deeplearningbook.org
15. Hesse, N., Bodensteiner, C., Arens, M., Hofmann, U., Weinberger, R., Schroeder, S.: An empirical study towards understanding how deep convolutional nets recognize falls. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018 Workshops. LNCS, vol. 11134, pp. 112–127. Springer, Cham (2019)
16. Ivorra, E., Ortega, M., Alcañiz, M., Garcia-Aracil, N.: Multimodal computer vision framework for human assistive robotics. In: 2018 Workshop on Metrology for Industry 4.0 and IoT, pp. 1–5. IEEE (2018)

17. Johnson, J., et al.: Inferring and executing programs for visual reasoning. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3008–3017, October 2017. https://doi.org/10.1109/ICCV.2017.325
18. Katircioglu, I., Tekin, B., Salzmann, M., Lepetit, V., Fua, P.: Learning latent representations of 3D human pose with deep neural networks. Int. J. Comput. Vis. (2018). https://doi.org/10.1007/s11263-018-1066-6
19. Kawana, Y., Ukita, N., Huang, J.B., Yang, M.H.: Ensemble convolutional neural networks for pose estimation. Comput. Vis. Image Underst. **169**, 62–74 (2018). https://doi.org/10.1016/j.cviu.2017.12.005
20. Leo, M., Medioni, G., Trivedi, M., Kanade, T., Farinella, G.: Computer vision for assistive technologies. Comput. Vis. Image Underst. **154**(Suppl. C), 1–15 (2017)
21. Leo, M., Del Coco, M., Carcagnì, P., Mazzeo, P.L., Spagnolo, P., Distante, C.: A technological framework to support standardized protocols for the diagnosis and assessment of ASD. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 269–284. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_19
22. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
23. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing **234**, 11–26 (2017). https://doi.org/10.1016/j.neucom.2016.12.038. http://www.sciencedirect.com/science/article/pii/S0925231216315533
24. Mahmud, T., Hasan, M., Roy-Chowdhury, A.K.: Joint prediction of activity labels and starting times in untrimmed videos. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5784–5793 (2017)
25. Nair, V., Budhai, M., Olmschenk, G., Seiple, W.H., Zhu, Z.: ASSIST: personalized indoor navigation via multimodal sensors and high-level semantic information. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018 Workshops. LNCS, vol. 11134, pp. 128–143. Springer, Cham (2019)
26. Nouredanesh, M., Li, A.W., Godfrey, A., Hoey, J., Tung, J.: Chasing feet in the wild: a proposed egocentric motion-aware gait assessment tool. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018 Workshops. LNCS, vol. 11134, pp. 176–192. Springer, Cham (2019)
27. Ortis, A., Farinella, G.M., D'Amico, V., Addesso, L., Torrisi, G., Battiato, S.: Organizing egocentric videos of daily living activities. Pattern Recogn. **72**, 207–218 (2017)
28. Park, H.S., Hwang, J.J., Niu, Y., Shi, J.: Egocentric future localization. In: CVPR 2016, pp. 4697–4705 (2016)
29. Perrett, T., Damen, D.: Recurrent assistance: cross-dataset training of LSTMs on kitchen tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1354–1362 (2017)
30. Pirri, F., Mauro, L., Alati, E., Sanzari, M., Ntouskos, V.: Deep execution monitor for robot assistive tasks. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018 Workshops. LNCS, vol. 11134, pp. 158–175. Springer, Cham (2019)
31. Ravì, D., et al.: Deep learning for health informatics. IEEE J. Biomed. Health Inform. **21**, 4–21 (2017)
32. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. arXiv preprint (2017)
33. Rhinehart, N., Kitani, K.M.: First-person activity forecasting with online inverse reinforcement learning. In: ICCV (2017)
34. Rodriguez, C., Fernando, B., Li, H.: Action anticipation by predicting future dynamic images. arXiv preprint arXiv:1808.00141 (2018)

35. Sawatzky, J., Gall, J.: Adaptive binarization for weakly supervised affordance segmentation. arXiv preprint arXiv:1707.02850 (2017)
36. Schydlo, P., Rakovic, M., Jamone, L., Santos-Victor, J.: Anticipation in Human-Robot Cooperation: A Recurrent Neural Network Approach for Multiple Action Sequences Prediction. arXiv e-prints, February 2018
37. Sciortino, G., Farinella, G.M., Battiato, S., Leo, M., Distante, C.: On the estimation of children's poses. In: Battiato, S., Gallo, G., Schettini, R., Stanco, F. (eds.) ICIAP 2017. LNCS, vol. 10485, pp. 410–421. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68548-9_38
38. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 806–813 (2014)
39. Soran, B., Farhadi, A., Shapiro, L.: Generating notifications for missing actions: don't forget to turn the lights off! In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4669–4677 (2016)
40. Soran, B., Lowes, L., Steele, K.M.: Evaluation of infants with spinal muscular atrophy type-I using convolutional neural networks. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 495–507. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_34
41. Sun, C., Shrivastava, A., Vondrick, C., Murphy, K., Sukthankar, R., Schmid, C.: Actor-centric relation network. arXiv preprint arXiv:1807.10982 (2018)
42. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: MovieQA: understanding stories in movies through question-answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4631–4640 (2016)
43. Toshev, A., Szegedy, C.: DeepPose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660 (2014)
44. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 98–106 (2016)
45. Wang, A., Dantcheva, A., Broutart, J.C., Robert, P., Bremond, F., Bilinski, P.: Comparing methods for assessment of facial dynamics in patients with major neurocognitive disorders. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018 Workshops. LNCS, vol. 11134, pp. 144–157. Springer, Cham (2019)
46. Wang, L., Wang, Z., Qiao, Y., Van Gool, L.: Transferring deep object and scene representations for event recognition in still images. Int. J. Comput. Vis. **126**(2), 390–409 (2018). https://doi.org/10.1007/s11263-017-1043-5
47. Yan, Z.: Computer vision for medical infant motion analysis: state of the art and RGB-D data set. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018 Workshops. LNCS, vol. 11134, pp. 32–49. Springer, Cham (2019)
48. Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D.: Pythia v0.1: the winning entry to the VQA challenge 2018. arXiv preprint arXiv:1807.09956 (2018)
49. Zhang, M., Ma, K.T., Lim, J.H., Zhao, Q., Feng, J.: Deep future gaze: gaze anticipation on egocentric videos using adversarial networks. In: Conference on Computer Vision and Pattern Recognition, pp. 4372–4381 (2017)
50. Zhang, Q., Yang, L.T., Chen, Z., Li, P.: A survey on deep learning for big data. Inf. Fusion **42**(Suppl. C), 146–157 (2018)

51. Zhang, Y., et al.: Physically-based rendering for indoor scene understanding using convolutional neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5057–5065. IEEE (2017)
52. Zhao, Z.Q., Zheng, P., Xu, S., Wu, X.: Object Detection with Deep Learning: A Review. arXiv e-prints, July 2018
53. Zhu, Y., Jiang, S.: Deep structured learning for visual relationship detection. In: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-2018) (2018)