Robust, Unsupervised, and Ubiquitous Pose and Object Information for Robot Action Recognition

Dan Scarafoni[1], Irfan Essa[1], David Kent[1], Harish Ravichandar[1], Sonia Chernova[1] and Thomas Plötz[1]

***Abstract*— Seamless human-robot collaboration requires accurate real-time activity recognition. For this task, human pose and object-interaction information provide essential context for classifying actions. Unfortunately, high fidelity pose information cannot be derived from RGB video, which constitutes the majority of the world's available data, and object recognition requires costly annotation. We present a technique to extract and utilize object features with no supervision and reliably use pose information from RGB video data. This technique, the Three-Dimensional Spatio-temporal Attention Mechanisms (3DSAM), innovatively represents pose and object information as evolutions of relevant pixel regions over time. 3DSAM derives contextually important areas from a video, extracts spatio-temporal features with a 3D convolutional neural network, then leverages a novel soft attention mechanism to enhance real-time, fine-grained action recognition essential for human-robot collaboration. We evaluate our approach on challenging datasets, as well as a specific case study in human-robot collaboration. We demonstrate that 3DSAM can achieve state-of-the-art RGB-based classification on scenarios essential for action recognition for human-robot collaboration.**

I. Introduction

Real-time activity recognition is an essential component of human-robot collaboration (HRC) as it allows the robot to see and understand human behavior. This allows the robot to safely and effectively engage with the human [1], [2], [3]. Human-robot collaboration frequently involves the analysis of fine-grained actions, which differ from one another subtly.

Previous works rely on using high-fidelity pose information [4], [5], requiring camera hardware which can capture depth information. The majority of pose-based techniques are not usable in modalities such as RGB, where pose-extraction techniques are not high enough quality to work. There is a real need to be able to utilize this modality, however, as RGB cameras do not have the distance limitation of RGB-D cameras, and are more widespread (particularly in factories and commercial buildings). Leveraging RGB systems would allow robotics techniques to integrate with existing infrastructure and video data. Indeed, this problem was recently highlighted as an important problem in current computer vision [6]. To compound this problem further, even with high-fidelity pose extraction, joint location error can still hinder classifiers [7], [8].

A similar hindrance exists with object-interaction information. Object detectors typically require supervised labeling [8], [9], [10], which makes these methods difficult to implement, as changing to a new problem or domain requires additional labeling.

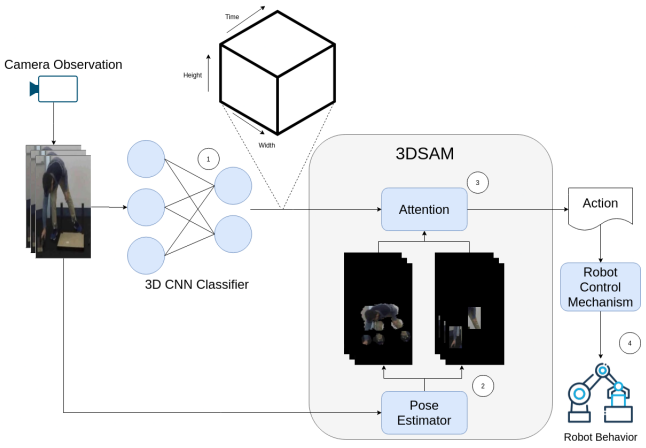


Fig. 1. The overall pipeline of our approach and example use case. By exploiting the spatio-temporal properties of 3D CNN feature maps (1), we can derive pose and object information (2), to form a novel attention mechanism (3) which emphasizes crucial feature maps to enhance fine-grained action recognition. This can then used, for example, by a control mechanism to make changes to the robot behavior in response to the human (4).

To address the need for robust, unsupervised pose and object-based classification in HRC, we present 3DSAM. 3DSAM breaks video images into separate, contextually important video streams. From these streams, 3DSAM extracts spatio-temporal features with a 3D convolutional neural network (3D CNN), preserving space and time relationships between feature maps. These are used to build a novel soft attention mechanism that uses this information to enhance classification. The end result is a system that can utilize error-prone pose information, extract and leverage unsupervised object information, and enhance theoretically any state-of-the-art 3D CNN action classifier.

Our specific contributions are:

1) We introduce a means of reliably utilizing error-prone pose information derived from RGB video data for action recognition.
2) We demonstrate a novel means of extracting and utilizing object features with no supervision.
3) We introduce a new representation for human pose and object information by tracking pixels relevant to that information over time.
4) We present a novel soft attention mechanism that utilizes spatio-temporal 3D CNN features, object in-

[1]Dan Scarafoni, Irfan Essa, David Kent, Harish Ravichandar, Sonia Chernova, and Thomas Plötz are with the Georgia Institute of Technology, School of Interactive Computing, North Ave NW, Atlanta, GA 30332. {danscarafoni,tploetz,irfan,dekent,harish.ravichandar,chernova}@gatech.edu

teraction data, and pose information to enhance fine-grained action classification on theoretically any 3D CNN action classifier.

5) We demonstrate a system that improves state-of-the-art classifiers on fine-grained action recognition datasets.

II. Related Work

A. Computer Vision for Fine-Grained, HRC-centered Action Recognition

Koppula et al. utilized RGB-D camera information to obtain object features and affordances to better plan actions [11]. Much work has used tagged objects to detect object interaction [12], [11], [13]. Mohseni-Kabir et al. utilized an RGB-D camera and an IMU to learn human activity primitives and objects [14]. Due to a lack of robustness, RGB data alone is generally not used to derive pose information. It is more common to use depth information or motion tracking for human body estimation. Physical markers, depth information, and supervised methods [15], [11] are more commonly used for object recognition.

There are several recent advancements in computer vision that are crucial to action recognition for human-robot collaboration. Action recognition for human-robot collaboration requires on fine-grained action recognition, where the differences between actions can differ only subtly [16]. Soft attention mechanisms, which draw attention to features in a neural network, have shown to be particularly important for fine-grained action recognition [17], [18], [19], [20]. Also, recent evidence indicates that minute, frame-by-frame differences are important for classifying fine-grained actions in particular [21]. Systems that preserve these spatio-temporal relations will be more discriminative than those that don't. 3D CNN feature maps preserve these relationships, as opposed to techniques such as LRCN, which convert frame-features into one-dimensional vectors before analyzing information temporally [22].

B. Pose and Object-Interaction Based Activity Recognition

Pose-based activity recognition is a well-established field in machine learning. Previous work has broken pose information down into sub-components to differentiate fine details of actions [23], [24]. Utilizing pose information, however, is challenging where the pose estimations are prone to error. This is particularly common in RGB data. Indeed, figuring out how to utilize pose (and other modality) information in RGB data was recently labeled as a key problem for computer vision research [6]. Recent papers have explored the utility of pose estimation maps, rather than the pose skeleton itself [25].

Object information is also very relevant for action recognition. Wei et al. trained an object detector, and used object interaction (defined in terms of object proximity) as a means of action classification [8].

3DSAM harmonizes these recent advancements and leverages them to build a novel pose and object-based action recognition system. 3DSAM leverages a unique, 3D CNN-based soft attention mechanism that preserves spatio-temporal relationships between feature maps to successfully leverage RGB-based pose and unsupervised object features. Unlike previous systems, our system is modular (usable with any 3DCNN), robust, and requires neither depth information nor object labeling.

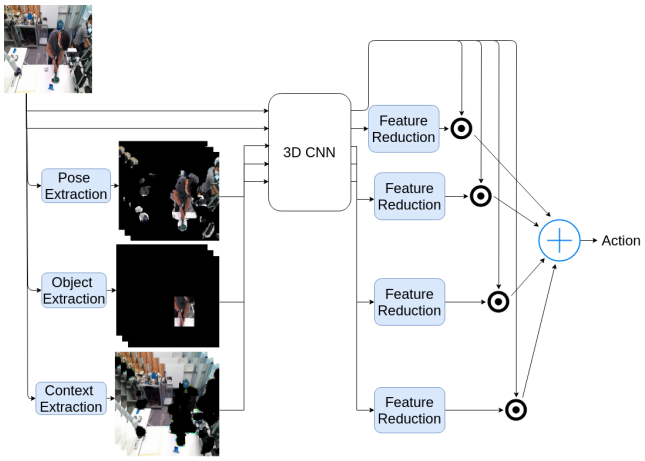


Fig. 2. A detailed view of the 3DSAM approach. Our approach creates pose, object, and context video streams from input video data, extracts 3D CNN features, and creates attention masks. It then utilizes these to draw attention to salient parts of input stream to enhance classification.

III. 3DSAM- Augmenting CNNs with RGB Pose and Unsupervised Object Features

A. Approach Description

Our 3DSAM approach creates "streams" for pose, object, and miscellaneous context data. These streams are modifications of the RGB video in which all pixel values are reduced to zero except those of semantic importance. For example, the pose stream is the video with all the pixels blocked out except those near the subject's joints. The object stream is the video with all pixels blocked out except those near the subject's hands.

From these (as well as the original RGB video) 3DSAM extracts 3D CNN features and derives attention weight tensors from each. The features of a 3D CNN are 4-dimensional tensors, preserving the relationships between height, width, and time of the pixels in the input stream. This we contrast with other feature extraction techniques that transform a frame or video sequence into a single one-dimensional vector, thus destroying explicit spatio-temporal relationships [26].

The attention mechanism point-wise multiplies the features from the RGB video by each of these weight tensors, emphasizing important features for classification. The explicit usage of separate pose and object modalities ensures that 3DSAM learns crucial information for planning motion around the human collaborator. The preservation of spatio-temporal features in the feature mapping also preserves the minute changes in space and time that are crucial to differentiating fine-grained actions for HRC [16]. A full overview of our approach is given in Fig. 2.

B. Pose Stream

Pose information is encoded in a spatio-temporal stream of the subject's skeletal joints. This is derived using convolutional pose machines [5]. Our model assumes a pose detection approach which, given a $3 \times H \times W$ RGB image i from a video I, generates an $H \times W$ mask image ℓ_{Pose} which is part of a larger sequence L_{Pose} of equal length to I. Every element in ℓ_{Pose} is either 1 if the corresponding pixel in i is part of the actor's body and 0 otherwise. The pose stream S_{pose} is a stream of RGB frames produced by point-wise multiplying this mask by the RGB image:

$$S_{pose} = I \odot L_{pose} \tag{1}$$

This produces a "tube" like structure if seen in three dimensions which illustrates the joint movements over time. Visualization of these streams is given in Fig. **??**. The output is then fed into the 3D CNN f to produce network features:

$$N_{pose} = f(S_{pose}) \tag{2}$$

N_{pose} is still a 4 dimensional tensor. As a final preparation step, the entire results are multiplied by a learned weight vector (implemented as a 3D convolution) to emphasize different features:

$$X_{pose} = N_{pose} \times W_{pose} \tag{3}$$

although this does not change the size of the output in space or time, this weight vector reduces the dimensionality of the stream of our network.

C. Object Stream

We note that in human-robot collaboration tasks, "object interaction" frequently involves holding the object in the hand. This is particularly true for industrial tasks, in which tool usage with, e.g., feet are rare. "Attaching shelf" is very similar to "attach table leg," except for the object. Because manual manipulation plays such a central role in object usage, we argue that information near the human hand can serve as an approximation for "object interaction." As such, we can gain insight into object usage by examining the hands of the person. Here, the hand locations are derived from the previously mentioned pose estimation method.

The object stream is thus derived from I by creating a mask L_{object} made of frames ℓ_{object} in which each element is 1 if the corresponding pixel is within 40 pixels $L1$ distance from the estimated hand joint in the frame and 0 otherwise: This empirically derived distance allows for a clear view of the subject's hands and the surrounding area. This value can be applied to any video stream where the subject is in full, clear view of the camera. The features for this stream are calculated in the same way as the pose stream.

$$S_{object} = I \odot L_{object} \tag{4}$$

$$N_{object} = f(S_{object}) \tag{5}$$

$$X_{object} = N_{object} \times W_{object} \tag{6}$$

D. Context Stream

We define contextual information as information in the scene that is not related to the pose or the object interaction. For example, differentiating between "attach table leg" and "attach chair leg" requires knowledge of the overall scene to differentiate the actions. Here, context information is derived by taking the input video and removing the pose information, i.e. $L_{context}$ is all the elements of I with the only the pose pixels blocked out:

$$L_{context} = \neg L_{pose} \tag{7}$$

$$S_{context} = I \odot L_{context} \tag{8}$$

$$N_{context} = f(S_{context}) \tag{9}$$

$$X_{context} = N_{context} \times W_{context} \tag{10}$$

E. RGB Stream

The above streams capture salient parts of the video stream, though they cannot capture important motifs that span across streams. For example, in an industrial task, a person may be carrying a large piece of wood, which could potentially span the entire image, and not localize cleanly into any of the above streams. Research indicates that the analysis of image sub-components captures different information than the whole [23]. Thus, in addition to the previous weights from separated streams, we also include attention weights derived from the full RGB stream:

$$N_{rgb} = f(I) \tag{11}$$

$$X_{rgb} = N_{rgb} \times W_{rgb} \tag{12}$$

F. The Convolutional Attention Mechanism

After all four streams are derived, they are concatenated with one another. The result is a tensor of shape:

$$X = C \times frames \times height \times width \tag{13}$$

where C is the number of features. The original RGB stream is then multiplied by each of these attention matrices $X^{(i)}$ to produce a new vector X_i' :

$$X'^{(i)} = I \odot X^{(i)} \tag{14}$$

These features are then passed through a global average pooling layer, and finally a fully-connected layer to obtain the final per-frame logits. The overall layout of our approach can be seen in Fig. 2.

IV. Experimental Setup

A. Datasets

We evaluate our approach on two datasets for fine-grained activity recognition: the IKEA Furniture Assembly dataset and MSR-DailyActivity3D [27], [28]. We also evaluate our system with a human-robot collaboration case study where a human and a Jaco robot work together to assemble a set of power drills. Several sample images from different streams can be seen in Fig. 3.

Although the pose information comes with each dataset, we extract pose information manually from the RGB data with convolutional pose machines to demonstrate our system can work with any dataset [5].

B. Modeling Parameters and Baseline Comparison

Our system is designed to be an add-on to existing RGB action recognition systems. 3DSAM is meant to provide a way to enhance the classification accuracy of 3D CNNs by exploiting pose and object information. To evaluate 3DSAM,

Approach	Accuracy
pose-based [27]	69% ± 0
C3D Network	55% ± 9
C3D + 3DSAM	**69% ± 3**
I3D Network	71% ± 2
I3D + 3DSAM	**75% ± 4**

TABLE I

PERFORMANCE OF THE I3D, 3DSAM, AND BASELINE POSE-BASED CLASSIFIER ON THE IKEA ASSEMBLY DATASET

Approach	Accuracy
pose-based [27]	69%
C3D Network	55%
C3D + 3DSAM	**69%**
I3D Network	71%
I3D + 3DSAM	**75%**

TABLE II

PERFORMANCE OF THE I3D, 3DSAM, AND BASELINE POSE-BASED CLASSIFIER ON THE IKEA ASSEMBLY DATASET

therefore, we evaluate its ability to increase state-of-the-art 3D CNN performance on fine-grained datasets. We compare it against the I3D network, a ubiquitously used state-of-the-art 3D CNN for action recognition [29], [21], [30]. We utilize weights pre-trained on ImageNet, Kinetics, and with the last layer retrained on our specific datasets [31]. To further test generalizability, we also evaluate performance on the C3D system, which was the most advanced 3D CNN for action recognition before I3D [32], [33]. For this system, we use weights pre-trained on the Sports 1-Million dataset [34], [35].

For the IKEA dataset and the case study, we use per-frame average precision over each class as the evaluation metric. For the DailyActivity3D dataset, we take 5 evenly spaced clips from each video, run them through the network, then average the end logits to get the final classification.

V. RESULTS

The results for the experiments on the IKEA dataset are shown in Tab. II We note that the I3D essentially learns a binary classifier- "spin" vs. "not spin." 3DSAM, however, can detect more classes. The IKEA dataset has an imbalance of classes, with nearly three of four frames being one class (class 3: "spin/unspin table leg"). Both systems tend to overemphasize this class as a result.

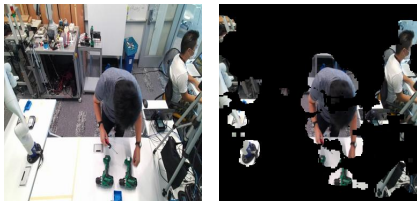
Similarly, we evaluate our approach on the MSR-DailyActivity3D dataset. The results are shown in Tab. III.

We evaluate the case study in the same manner as the IKEA dataset, though only training the network once on the data. We note that the 3DSAM improves considerably over the baseline, increasing per-frame accuracy by nearly 40%. Results are shown in Tab. IV.

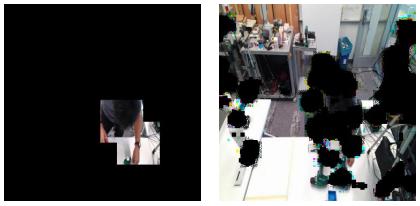
Approach	Accuracy
C3D Network	22%
C3D + 3DSAM	**29%**
I3D Network	58%
I3D + 3DSAM	**62%**

TABLE III

PERFORMANCE ON THE MSR-DAILYACTIVITY3D DATASET



(a) (b)



(c) (d)

Fig. 3. Sample stream frames from the case study showing stream analysis of a human working with a Jaco arm to assemble a power drill.

Approach	Accuracy
I3D Network	43%
I3D + 3DSAM	**83%**

TABLE IV

PERFORMANCE ON THE DRILL ASSEMBLY CASE-STUDY.

VI. CONCLUSION

We presented a method for augmenting real-time state-of-the-art classifiers for computer vision-based activity recognition, with particular importance to HRC applications. By separating input video into component streams, we show a novel way of augmenting fine-grained action recognition by explicitly leveraging pose and object features and soft attention mechanisms. Moreover, we presented a solution to a critical problem in the computer vision and HRI communities: the utilization of pose and object modalities in RGB videos. Our system brings action recognition techniques to the vast majority of the world's data and allows integration with the world's existing camera infrastructure.

REFERENCES

[1] S. V. Albrecht and P. Stone, "Autonomous agents modelling other agents: A comprehensive survey and open problems," *Artificial Intelligence*, vol. 258, pp. 66–95, May 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370218300249

[2] J. Heard, R. Heald, C. E. Harriott, and J. A. Adams, "A Diagnostic Human Workload Assessment Algorithm for Human-Robot Teams," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. Chicago IL USA: ACM, Mar. 2018, pp. 123–124. [Online]. Available: https://dl.acm.org/doi/10.1145/3173386.3176983

[3] S. M. al Mahi, M. Atkins, and C. Crick, "Learning to Assess the Cognitive Capacity of Human Partners," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. Vienna Austria: ACM, Mar. 2017, pp. 63–64. [Online]. Available: https://dl.acm.org/doi/10.1145/3029798.3038430

[4] F. Baradel, C. Wolf, and J. Mille, "Human Action Recognition: Pose-Based Attention Draws Focus to Hands," *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, vol. 2018-Janua, pp. 604–613, 2018.

[5] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 4724–4732, 2016.

[6] H. B. Zhang, Y. X. Zhang, B. Zhong, Q. Lei, L. Yang, J. X. Du, and D. S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors (Switzerland)*, vol. 19, no. 5, pp. 1–20, 2019.

[7] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Pattern Recognition5*, vol. 171, no. April, pp. 118–139, 2018. [Online]. Available: https://doi.org/10.1016/j.cviu.2018.04.007

[8] P. Wei, Y. Zhao, N. Zheng, and S. C. Zhu, "Modeling 4D Human-Object Interactions for Joint Event Segmentation, Recognition, and Object Localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1165–1179, 2017.

[9] C. Wu, J. Zhang, O. Sener, B. Selman, S. Savarese, and A. Saxena, "Watch-n-Patch: Unsupervised Learning of Actions and Relations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 467–481, 2015.

[10] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[11] H. S. Koppula and A. Saxena, "Anticipating Human Activities Using Object Affordances for Reactive Robotic Response," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 14–29, 2016.

[12] G. Milliez, R. Lallement, M. Fiore, and R. Alami, "Using human knowledge awareness to adapt collaborative plan generation, explanation and monitoring," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2016-April, pp. 43–50, 2016.

[13] S. Devin and R. Alami, "An implemented theory of mind to improve human-robot shared plans execution," in *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2016-April. IEEE, 2016, pp. 319–326.

[14] A. Mohseni-Kabir, C. Li, V. Wu, D. Miller, B. Hylak, S. Chernova, D. Berenson, C. Sidner, and C. Rich, "Simultaneous learning of hierarchy and primitives for complex robot tasks," *Autonomous Robots*, vol. 43, no. 4, pp. 859–874, 2019. [Online]. Available: https://doi.org/10.1007/s10514-018-9749-y

[15] B. Reily, Q. Zhu, C. Reardon, and H. Zhang, "Simultaneous learning from human pose and object cues for real-time activity recognition," *arXiv preprint arXiv:2004.03453*, 2020.

[16] C. Lea, R. Vidal, and G. D. Hager, "Learning convolutional action primitives for fine-grained action recognition," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2016-June, pp. 1642–1649, 2016.

[17] N. Raman and S. J. Maybank, "Activity recognition using a supervised non-parametric hierarchical HMM," *Neurocomputing*, vol. 199, pp. 163–177, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2016.03.024

[18] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition," pp. 1227–1236, 2019. [Online]. Available: http://arxiv.org/abs/1902.09130

[19] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek, "VideoLSTM convolves, attends and flows for action recognition," *Computer Vision and Image Understanding*, vol. 166, pp. 41–50, 2016.

[20] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition,*, vol. 2017-Janua, no. typically 25, pp. 3165–3174, 2017.

[21] D. Dwibedi, P. Sermanet, and J. Tompson, "Temporal Reasoning in Videos Using Convolutional Gated Recurrent Units," in {*CVPR*} *Workshops*, 2018, pp. 1111–1116.

[22] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2015.

[23] G. Cheron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN Features for Action Recognition," in *ICCV*, 2015, p. 24.

[24] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 724–731.

[25] M. Liu and J. Yuan, "Recognizing Human Actions as the Evolution of Pose Estimation Maps," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1159–1168.

[26] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," pp. 1–10, 2013. [Online]. Available: http://arxiv.org/abs/1312.6199

[27] S. Toyer, A. Cherian, T. Han, and S. Gould, "Human Pose Forecasting via Deep Markov Models," in *DICTA 2017 - 2017 International Conference on Digital Image Computing: Techniques and Applications*, vol. 2017-Decem, 2017, pp. 1–8.

[28] J. Wang, Z. Liu, and Y. Wu, "Learning Actionlet Ensemble for 3D Human Action Recognition," *Computer Vision and Image Understanding5*, vol. 36, no. 9783319045603, pp. 11–40, 2014.

[29] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose MoTion Representation for Action Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7024–7033, 2018.

[30] I. Funke, S. T. Mees, J. Weitz, and S. Speidel, "Video-based surgical skill assessment using 3D convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 7, pp. 1217–1225, 2019.

[31] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," 2017. [Online]. Available: http://arxiv.org/abs/1705.07750

[32] D. Liu and T. Jiang, "Deep Reinforcement Learning for Surgical Gesture Segmentation and Classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11073 LNCS, pp. 247–255, 2018.

[33] T. Mahmud, M. Hasan, and A. K. Roy-Chowdhury, "Joint prediction of activity labels and starting times in untrimmed videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5773–5782.

[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," 2014. [Online]. Available: http://arxiv.org/abs/1412.0767

[35] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1725–1732, 2014. [Online]. Available: http://www-cs.stanford.edu/groups/vision/pdf/karpathy14.pdf{\%}5Cnhttp://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909619{\%}0Apapers3://publication/doi/10.1109/CVPR.2014.223