EFFICIENT HUMAN ACTIVITY CLASSIFICATION FROM EGOCENTRIC VIDEOS INCORPORATING ACTOR-CRITIC REINFORCEMENT LEARNING

Yantao Lu, Yilan Li and Senem Velipasalar
Department of Electrical Engineering and Computer Science
Syracuse University, Syracuse, NY, USA 13244-1240

ABSTRACT

In this paper, we introduce a novel framework to significantly reduce the computational cost of human temporal activity recognition from egocentric videos while maintaining the accuracy at the same level. We propose to apply the actor-critic model of reinforcement learning to optical flow data to locate a bounding box around region of interest, which is then used for clipping a sub-image from a video frame. We also propose to use one shallow and one deeper 3D convolutional neural network to process the original image and the clipped image region, respectively. We compared our proposed method with another approach using 3D convolutional networks on the recently released Dataset of Multimodal Semantic Egocentric Video. Experimental results show that the proposed method reduces the processing time by 36.4% while providing comparable accuracy at the same time.

Index Terms— activity classification, reinforcement learning, actor critic

1. INTRODUCTION

There have been many methods for human activity classification, which rely on third-person video data [1, 2, 3, 4, 5] from static cameras watching activities of person(s). Compared to human activity video datasets obtained from static cameras, there has been much less video data from egocentric cameras. Similarly, compared to works that use static cameras installed in the environment, there has been relatively less work using egocentric videos, meaning providing the first-person view from wearable cameras.

Heilbron et al. [6] presented the ActivityNet, which is a large-scale video benchmark for human activity understanding, and proposed a method based on 3D Convolutional Neural Networks (CNNs). In this video dataset, majority of videos are not egocentric. Karpathy et al. [1] proposed a method for large-scale video classification, and presented results on the UCF-101 Action Recognition Dataset [7], which mostly contains third-person videos. Instead of using 2D CNN and LSTM, different approaches have been presented using 3D CNNs for learning spatiotemporal features [3]. Montes et al. [4] use a 3D CNN together with LSTM to achieve temporal activity detection in untrimmed videos. Instead of using LSTM, Buch et al. [5] use 3D CNN together with Gated Recurrent Units on videos from a third-person view.

Many approaches have been developed, which employ deep neural networks to perform human activity classification. In addition to the networks having deeper structures, higher resolution image data needs to be processed in many cases. This increases the computational complexity. Thus, researchers have also focused on speeding up the processing [8, 9, 10, 11]. However, these models are mostly tailored to particular network structures, and may not generalize well to new architectures. Minh et al. [12] introduced a recurrent neural network-based model to represent visual attention, and applied it to the image classification task and a simple game.

Reinforcement learning (RL) provides a mathematical framework for learning or deriving policies that map situations (i.e. states) into actions with the goal of maximizing an accumulative reward [13]. Unlike supervised learning, in RL the *agent* (i.e. learner) learns the *policy* for decision making through interactions with the *environment*. The goal of the agent is to maximize the cumulative *reward* by taking the optimal action at each time step according to the current *state* while considering the trade-off between explorations and exploitations. The combination of conventional Q-learning and deep neural network, i.e. Deep Q-network (DQN) [14], provides a breakthrough in deep reinforcement learning (DRL). However, the neural network in DQN needs to accumulate enough samples of values, and the data needed for its training can either come from a model-based simulation or from actual measurement [15]. Originally developed by DeepMind, the DRL provides a promising data-driven, adaptive technique in handling large state space of complicated control problems [16]. The actor-critic deep reinforcement learning [17] has overcome difficulties in learning control policies of systems with continuous state and action space, which provides a potential solution for efficient real-time processing of video clips in our case.

In this paper, we propose a novel approach to significantly reduce the computational cost of human activity classification from egocentric videos while maintaining the accuracy at the same level. We leverage the actor-critic model of RL, and apply it to optical flow data to determine how to move a bound-

The information, data, or work presented herein was funded in part by National Science Foundation under Grants 1739748 and 1816732, and by the Advanced Research Projects Agency-Energy, U.S. Department of Energy, under Award Number DE-AR0000940. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ing box in x and y directions to maximize the reward, and find an optimal region of interest. The bounding box is used for clipping a portion of the image. We also propose to use one shallow and one deeper convolutional neural network to process the original image and the clipped image region, respectively. This overall proposed architecture will henceforth be referred to as the Deep-Shallow Network. We compared our proposed method with another approach, using 3D convolutional networks for activity recognition, on the recently released Dataset of Multimodal Semantic Egocentric Video. The results will be presented in Sec. 3.

2. PROPOSED METHOD

The overall Deep-Shallow Network, shown in Fig. 1, is composed of a shallow network, a deeper network and an image clipper. Both shallow and deep feature extractors are 3D convolutional neural networks (CNNs). The shallow feature extractor takes the original images as input, and uses relatively larger kernels and fewer layers to extract environment features from the larger original image. On the other hand, the deep feature extractor takes the clipped image regions as input, and uses smaller kernels to extract activity features. The details of the shallow and deep network models can be seen in Fig. 2.

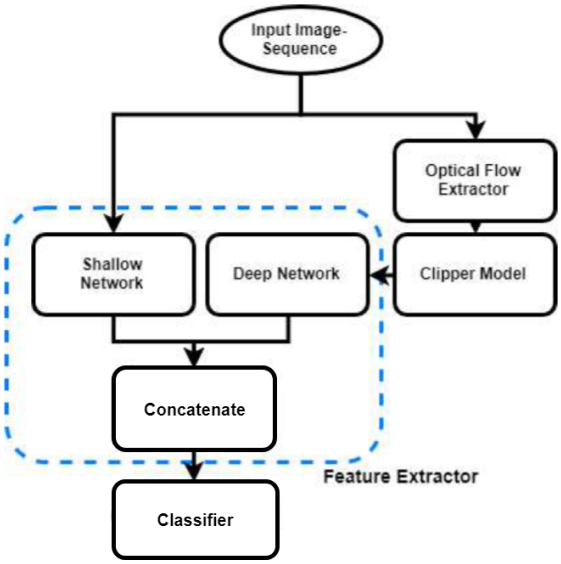

Fig. 1. Overall structure of the proposed Deep-Shallow Network.

The image clipper is trained based on the actor-critic model of RL. The input of the actor-critic model is the optical flow data extracted from the original images. Extracted features from the shallow and deep networks are concatenated, and followed by fully connected layers to obtain classification results.

As a result of reducing the complexity of the network structure, and only processing the regions of interest with a deep network, the proposed Deep-Shallow Network can significantly increase the processing speed, while maintaining the same level of accuracy with a state-of-the-art 3D CNN network.

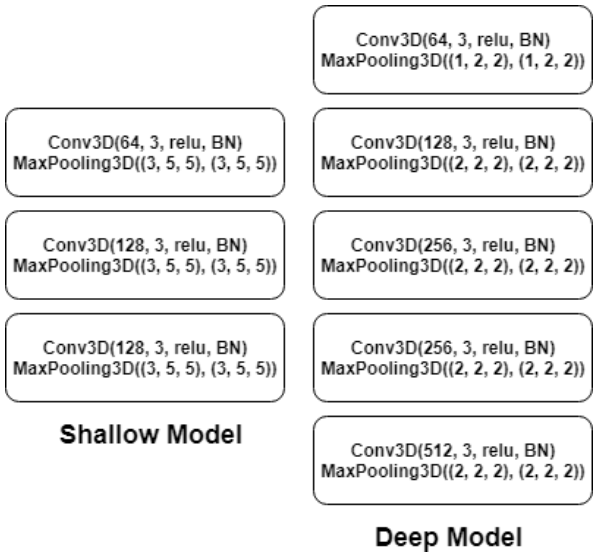

Fig. 2. Deep and shallow network model details.

2.1. Clipper Model Trained with Deep Reinforcement Learning

The main goal of the proposed approach is to reduce the computational complexity without sacrificing accuracy. A deep reinforcement learning (DRL)-based approach is adopted in this work to train the image clipper, which determines the location of the region of interest by moving a fixed-size bounding box. The height and width of the bounding box is half of the original image size. We build a standard reinforcement learning [16] setup up to derive the correlation between each state-action pair (s, a) of the system under control and its *value function* $Q(s, a)$ in discrete decision epochs. At each decision epoch t_k of the processing, the agent, i.e. the video frame at t_k, is at state s_k, and performs inference using deep neural network to select action a_k according to the policy π. We define the control actions as (Δ_x, Δ_y) with real values, which represent offsets of the bounding box in x and y directions, respectively. Since our problem has continuous output space, an actor-critic-based DRL [14] is adopted. Under a certain policy π, the value of $Q(s, a)$ estimates the accumulated discounted reward of each state-action pair:

$$Q(s, a) = \mathbf{E}(\Sigma_{k=0}^{\infty} \lambda^k r_k(s_k, a_k) | s_0 = s, a_0 = a)) \qquad (1)$$

where r_k is the total reward observed at decision epoch t_k. To accelerate learning, and avoid oscillations or divergence in the parameters, we employ an experience replay and target network [17]. The experience replay updates the weights of the target network θ' based on learned network weights θ by:

$$\theta' = \tau\theta + (1 - \tau)\theta', \qquad \tau \ll 1 \qquad (2)$$

The actor model is a feed-forward neural network composed of three fully-connected hidden layers with rectified

linear units (ReLU) as the activation function. It is used to predict the optimal action based on the current state $\mathbb{S}_t$. The number of neurons in fully connected hidden layers are 64, 128 and 128, respectively. The output layer size is 2 providing the horizontal and vertical offsets for the bounding box.

The critic model is another feed-forward neural network that evaluates the state and action pair, and the evaluation is used by the actor model to update its control policy in particular gradient direction. The critic model has two hidden layers. The first layer contains two separate fully-connected structures and the number of hidden neurons in each is 32. The addition of outputs from the first hidden layer is fed into the second layer which has 64 hidden neurons. The inputs of the critic model are $\mathbb{S}_t$ and $\mathbb{A}_t$, and the output is a single value $Q(\mathbb{S}_t, \mathbb{A}_t)$. The actor-critic framework is shown in Fig. 3.

During training, the actor model is trained using pair data $(\mathbb{S}_t, \mathbb{A}_t)$ to predict the optimal action $\mathbb{A}_t$ based on current agent state $\mathbb{S}_t$. Next agent state $\mathbb{S}_{t+1}$ is calculated through environment based on $\mathbb{A}_t$ and is used to predict optimal $\mathbb{A}_{t+1}$ by actor model. The critic model evaluates the resulting $\{\mathbb{S}_{t+1}, \mathbb{A}_{t+1}\}$ pair by predicting a Q-value to fine-tune action prediction. Therefore, the weights in actor model are updated by the gradient between actor and critic model, using chain rule $dQ/dW_{actor} = dQ/dW_{critic} \times dW_{critic}/dW_{actor}$. W_{actor} and W_{critic} indicate the weights of *actor* and *critic* models, respectively.

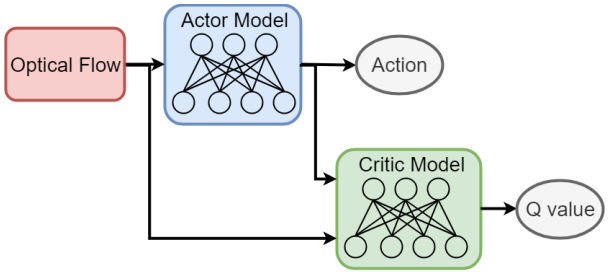

Fig. 3. Actor-critic based clipper model.

Example images from four different egocentric videos are shown in Fig. 4 together with the bounding boxes placed via the actor critic model. The first, second and third rows show frames $(t-10)$, t and $(t+10)$, respectively. As can be seen, the box placed by the actor-critic model moves inside the image to determine a focus region of interest.

3. EXPERIMENTAL RESULTS

We compared our proposed Deep-Shallow network with a commonly used 3D CNN [3], which will be referred to as C3D. We have used a recently released dataset called Dataset of Multimodal Semantic Egocentric Video (DoMSEV) [18]. DoMSEV contains 80 hours of egocentric video covering a wide range of activities. The videos were recorded using either a GoPro Hero camera or a built setup composed of a 3D Inertial Movement Unit (IMU) attached to the Intel Realsense R200 RGB-D camera. The activities performed while recording include walking, running, standing, browsing, driving, biking, eating, cooking, eating, observing, in conversation, playing, and shopping. We selected 11 videos (8 Tourism and 3 Daliy life videos), and five activities $(walking, running, standing, in\ conversation, browsing)$ as labels. We segment the videos into video clips of 60 frames with 50% overlapping. Then, we randomly separate 80% of data for training and 20% for testing. 20% of the training data is used for validation. The curves of training and validation loss of our Deep-Shallow model are shown in Fig. 5. The loss and reward curves of the actor-critic-based clipper model are shown in Fig. 6 and Fig. 7, respectively.

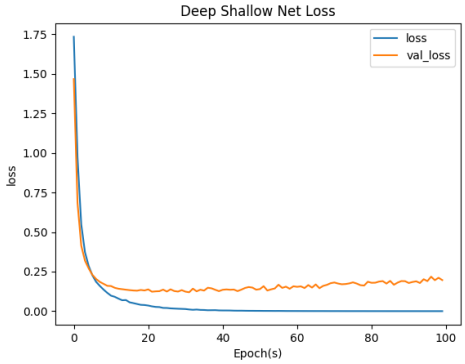

Fig. 5. Deep-Shallow network training loss.

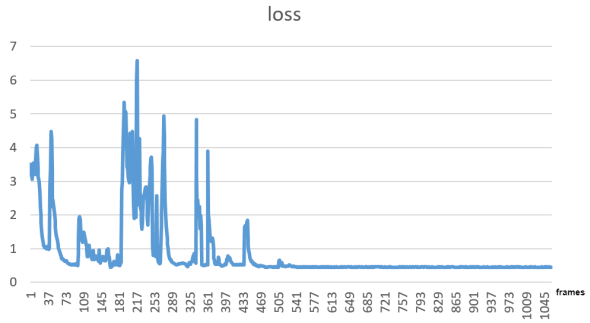

Fig. 6. Actor model training loss.

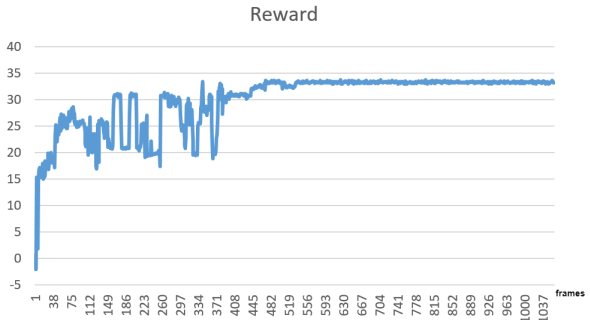

Fig. 7. Actor model training reward.

As mentioned above, we compared the performances of the proposed method and the traditional C3D in terms of speed and accuracy. For all the video clips (60 frame duration) in one video, we measured how long it takes to process them, and took the average. As shown in Table 1, the aver-

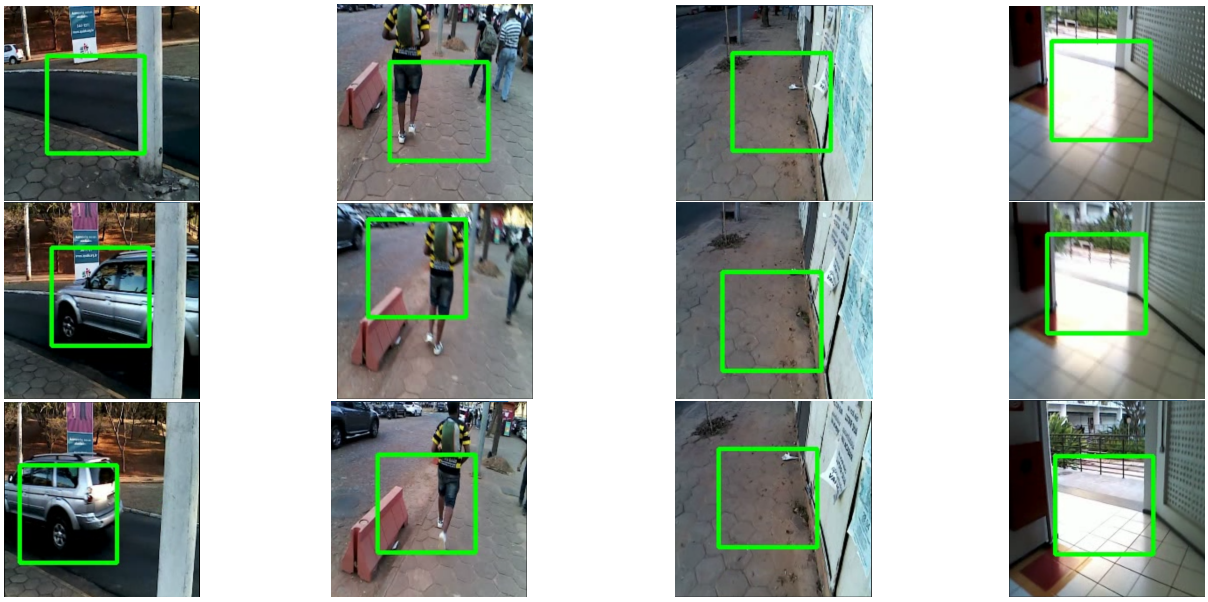

Fig. 4. Examples showing the autonomously placed bounding boxes. 1st, 2nd and 3rd rows show frames (t-10), t and (t+10), respectively.

age processing time per clip is 906 ms for C3D, while the average processing time per clip is 576 ms for the proposed Deep-Shallow model. In other words, our proposed model is 36.4% faster than the C3D as seen in Fig. 8.

The precision and recall values for each class are shown in Fig. 9 and Fig. 10, respectively. The average precision of the C3D and the proposed Deep-Shallow network are 0.72 and 0.71, respectively. The average recall of the C3D and the proposed method are 0.75 and 0.74, respectively. As seen in Table 1, the C3D and the proposed approach achieves 74% and 72.9% overall accuracy, respectively. In summary, the proposed approach provides a significant increase in processing speed with only 1.1% decrease in the accuracy.

	C3D	Deep-Shallow
Avg. process. time/clip	906 ms	576 ms
Overall accuracy	0.740	0.729

Table 1. Comparison table

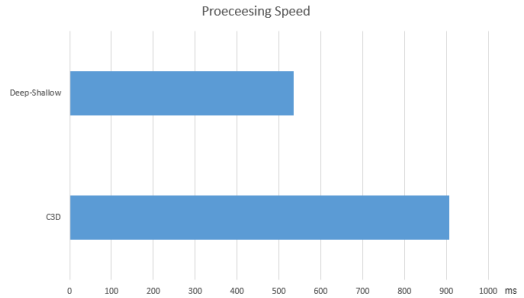

Fig. 8. Processing speed comparison

4. CONCLUSION

We have presented a novel method to efficiently perform human activity classification from egocentric videos by incorporating actor-critic model of reinforcement learning. Actor-critic reinforcement learning allows placing a bounding box on a region of interest, and clipping that region. Then, only the clipped region is processed through a deeper network, while the entire image is processed by a shallow one. This strategically reduced complexity of the network structure provides significant increase in the processing speed, while maintaining the same level of accuracy with a state-of-the-art 3D CNN network.

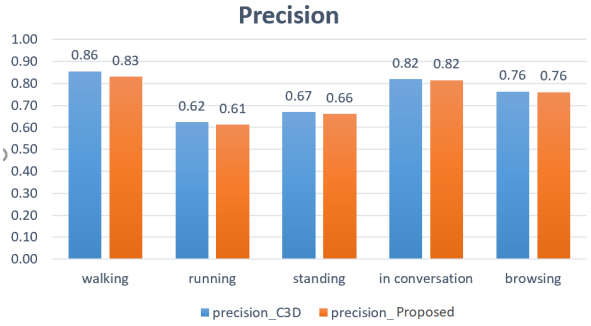

Fig. 9. Precision values for each activity class

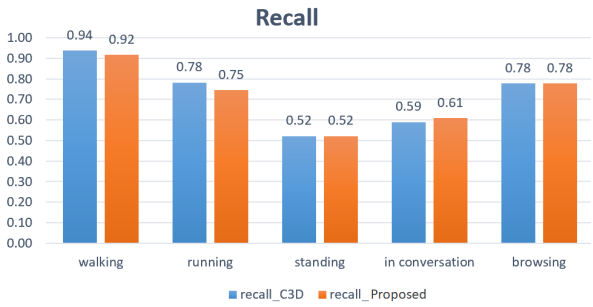

Fig. 10. Recall values for each activity class

5. REFERENCES

[1] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li, "Large-scale video classification with convolutional neural networks," in *Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[2] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Comp. Vision and Pattern Recogn.*, 2015, pp. 677–691.

[3] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Washington, DC, USA, 2015, ICCV '15, pp. 4489–4497, IEEE Computer Society.

[4] A. Montes, A. Salvador, S. Pascual, and X. Giroinieto, "Temporal activity detection in untrimmed videos with recurrent neural networks," *arXiv:1608.08128*, 2017.

[5] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles, "Sst: Single-stream temporal action proposals," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 6373–6382.

[6] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[7] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," 12 2012.

[8] Mikhail Figurnov, Aizhan Ibraimova, Dmitry P. Vetrov, and Pushmeet Kohli, "Perforatedcnns: Acceleration through elimination of redundant convolutions," in *NIPS*, 2016.

[9] Xiangyu Zhang, Jianhua Zou, Xiang Ming, Kaiming He, and Jian Sun, "Efficient and accurate approximations of nonlinear convolutional networks," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1984–1992, 2015.

[10] Emily Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA, 2014, NIPS'14, pp. 1269–1277, MIT Press.

[11] Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin, "Compression of deep convolutional neural networks for fast and low power mobile applications," *CoRR*, vol. abs/1511.06530, 2015.

[12] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu, "Recurrent models of visual attention," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, Cambridge, MA, USA, 2014, NIPS'14, pp. 2204–2212, MIT Press.

[13] Richard S Sutton and Andrew G Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al., "Human-level control through deep reinforcement learning," *Nature*, 2015.

[15] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al., "Mastering the game of go with deep neural networks and tree search," *nature*, 2016.

[16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[17] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[18] Michel Melo Silva, Washington L. S. Ramos, João P. K. Ferreira, Felipe C. Chamone, Mario F. M. Campos, and Erickson R. Nascimento, "A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, 2018, pp. 2383–2392.