딥러닝 기반 인간 동작 예측 기법 서베이

Matthew Marchellus, 박인규
인하대학교 정보통신공학과
{marchellusmatthew@gmail.com, pik@inha.ac.kr}

Human Motion Prediction with Deep Learning: A Survey

Matthew Marchellus, In Kyu Park
Department of Information and Communication Engineering, Inha University

요 약

인간 자세 추정 연구는 최근 크게 주목 받고 있는 연구 분야이다. 본 연구는 또한, 자기 지도 학습이라고 명명된 딥러닝 기법이 부상하면서 여러 문제가 해결되고 있다. 본 논문에서는, 이러한 문제를 해결하는 딥러닝 기반 인간 자세 추정 방법들을 유형별로 분류해본다. 그리고 각 분류별 설명과 함께 대표적인 방법들을 소개한다. 마지막으로, 결론에서는 본 연구가 앞으로 나아갈 방향에 대한 논의를 제시한다.

1. Introduction

The ability to anticipate subsequent human motion for machine is beneficial as it opens many possible applications. Developments of animation, VR and AR can be further improved with the existence of a flawless human motion prediction model. But more importantly, by inventing a model whose result is almost identical with human behavior implies that we have acquired knowledge on how human behaves in certain tasks. Such knowledge is helpful in applications such as sports, security, autonomous driving car, robotics, smart user interface and more.

Prediction task in general is challenging for machines but predicting subsequent action of a human is more complicated for machines. We humans can anticipate a set of actions from another human that is about to transpire when given some context or clues. The same cannot be said for machines. Even with deep learning tools, for every new context or clues we introduce that hints a future action, a modification to the model must be made for it to produce substantial result.

Attempts has been made to further advance machines capability to predict the future. Early on researchers attempted to utilize a recurrent neural network to extrapolate inputs into the future [1]. It was later proven that a simple RNN is insufficient, thus an additional module was introduced to further improve the result.

Many diverse works exist in the human motion prediction domain. We limit the scope of this survey to motion prediction methods using deep learning that takes in a sequence of data, which leaves aside methods that utilize only a single data as input (one-shot prediction). Moreover, we further limit the scope of the survey only to methods that generates a human motion prediction using a human body representation (3D human skeletons, 3D mesh, etc.).

This survey is organized as follows. First, Section 2 emphasizes what classifies as a human motion prediction and defines some key differences in other people's works. The subsection of Section 2 will explain those key

differences in detail. Afterwards we will highlight the datasets that are commonly used by human motion prediction methods in Section 3. Section 4 contains a discussion regarding possible future works that can be explored further. Finally, Section 5 summarizes the survey and draws conclusions about this work.

2. Human Motion Prediction

We define human motion prediction as the ability to predict the next action/motion from a human, given certain context which in this case is a sequence of data. Within this problem domain, there are various kinds of solutions proposed by researchers to address a variety of problems. In this section, we will compare some of those methods according to some key differences. This includes their network backbone, choice of human body representation, and their prediction type.

2.1. Network Backbone

Understanding details from a spatio-temporal representation of a data requires specialized network such as Recurrent models and Generative models. Throughout the years we see those methods gets increasingly intricate. Here we provide a brief explanation according to each network type that are present in the human motion prediction.

One of the first recurrent models is a combination of LSTMs and Encoder-Decoder architecture, called ERD, was proposed by Fragkiadaki et al. [1]. It produces a favorable result at that time, but it was later proved that it tends to converge to a mean position over time. A modification to the network called acLSTM was proposed which alternates the input between the ground truth data and the prediction result at previous timestep to the LSTM [3]. This reduces the error accumulation nature of RNN as it must predict using a degraded input (the networks output) while also forcing the network to be robust which results in the ability to predict long sequences.

Generative models have the advantage of not accumulating errors unlike recurrent models. We observe that an Autoencoder architecture can produce motion predictions [2]. By selectively connecting the joints of human body and grouping them by their respective limbs, they can produce a better prediction result (shown by their quantitative score) compared to ERD [1].

2.2. Human Body Representations

There are a few models to choose to represent the human body and use it to visualize the prediction result. Most early methods predict human motion over 3D keypoint in the Euclidean Space [1, 2, 3, 4, 6, 7]. When visualizing said model, the result is similar with human skeleton. This model is sufficient to show human motion, though it is very limited as a human body consists more of than just its joints.

It was not until a few years later where a full 3D body model was utilized in human motion prediction. Zhang et al. [5] proposed a prediction method that utilized the parametric body model SMPL [15]. The result is a 3D body model which contains more intricate detail compared to 3D keypoint. As a result, imperfections become more obvious to point out and it allows methods to use cost functions that were not available in 3D keypoint.

The model mentioned aforehand is the commonly used model though other models have been used in human motion prediction. Yuan et al. [9] proposed a Reinforcement Learning based human motion prediction method that uses a humanoid model in the physics simulation MuJoCo [16]. Yuan et al. [12] utilized 3D point cloud and predicts each point in the Euclidian Space.

2.3. Prediction Type

Prediction in this domain can be classified into 3 types. At the beginning prediction was done with simple regression method. It was discovered later that the multimodal nature of human motion requires more than regression and thus making the network produce a diverse result leads to better result. Finally, there is context-aware prediction that predicts human motion based on specific context given to the model.

Initially human motion prediction was attempted using simple regression concept. Networks in this prediction type can model simple human action (walking, running) but is

incapable of predicting complex human action such as eating, taking photo [14]. Moreover, networks in this category tends to converge to a mean position over time since most of these networks were not designed with the multimodal nature of human motion in mind.

Diverse prediction is mostly similar to regression-based prediction, though it is built with the multimodal nature of human motion in mind. Yan et al. [4] proposed a combined architecture with LSTM and VAE called MT-VAE. It managed to achieve diverse and plausible result. Afterwards, Yuan et al. [7] stated that previous methods focused more on the stochastic network while disregarding the sampling technique. They proposed novel sampling technique DLow to diversify the sampling process.

Context-aware prediction [10, 11] is different from the two mentioned before. Another input is given to the model alongside previous motion sequence. Said input can be an object or a person in the scene. As a result, the model will predict subsequent human motion depending on the object/person.

3. Dataset

Applicable datasets for human motion prediction mostly are video dataset. Most of the methods we present in this paper are highly dependent on the body type that is provided in the dataset. For example, the Human3.6M [14] dataset features RGB images, depth maps (time-of-flight range data), poses, scanned 3D surface meshes of all actors, silhouette masks, and 2D bounding boxes. As a result, the Human3.6M dataset can be used in many scenarios [1, 2, 4, 5, 6, 7, 8, 13]. However, to work with uncommon body model a specific dataset is required. For example, Yuan et al. [12] predicts motion using point cloud which utilized a dataset comprised of human scans.

4. Future Works

Recent works have produced a much better result when compared to earlier works, though many problems remain unexplored. For example, prediction on abstract human motion such as dancing have not been explored. Experiments could be done on networks that have not been utilized in human motion prediction. Furthermore, most current works predict only the big moving parts of the human body (head, limbs, and torso). Research can be done on methods that can predict the full extent of moving parts in human body (such as face and fingers) and is done simultaneously.

5. Conclusion

In this work, we present an analysis of the human motion prediction problem domain. We propose a classification of human motion prediction model based on deep learning. We explain the classes and provide an example for said classes. We then describe the required dataset to conduct research in this problem domain. Lastly, we present some ideas regarding possible future works in this field.

Acknowledgement

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFCIT1901-06. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2020-0-01389, Artificial Intelligence Convergence Research Center (Inha University))

References

[1] Fragkiadaki, K., Levine, S., Felsen, P., & Malik, J. (2015). Recurrent Network Models for Human Dynamics. 2015 IEEE International Conference on Computer Vision (ICCV), 4346-4354.

[2] Bütepage, J., Black, M.J., Kragic, D., & Kjellström, H. (2017). Deep Representation Learning for Human Motion Prediction and Classification. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1591-1599.

[3] Zhou, Y., Li, Z., Xiao, S., He, C., Huang, Z., & Li, H. (2018). Auto-Conditioned Recurrent Networks for Extended Complex Human Motion Synthesis. arXiv: Learning.

[4] Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., & Lee, H. (2018). MT-VAE: Learning Motion Transformations to Generate Multimodal Human Dynamics. ECCV.

[5] Zhang, J.Y., Felsen, P., Kanazawa, A., & Malik, J. (2019). Predicting 3D Human Dynamics From Video. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 7113-7122.

[6] Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., & Tian, Q. (2020). Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 211-220.

[7] Yuan, Y., & Kitani, K.M. (2020). DLow: Diversifying Latent Flows for Diverse Human Motion Prediction. ECCV.

[8] Zhang, Y., Black, M.J., & Tang, S. (2020). Perpetual Motion: Generating Unbounded Human Motion. ArXiv, abs/2007.13886.

[9] Yuan, Y., & Kitani, K. (2020). Residual Force Control for Agile Human Behavior Imitation and Extended Motion Synthesis. ArXiv, abs/2006.07364.

[10] Corona, E., Pumarola, A., Alenyà, G., & Moreno-Noguer, F. (2020). Context-Aware Human Motion Prediction. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6990-6999.

[11] Cao, Z., Gao, H., Mangalam, K., Cai, Q., Vo, M., & Malik, J. (2020). Long-term Human Motion Prediction with Scene Context. ECCV.

[12] Yuan, S., Li, X., Tzes, A., & Fang, Y. (2020). 3DMotion-Net: Learning Continuous Flow Function for 3D Motion Prediction. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 8154-8160.

[13] Zhang, Y., Black, M.J., & Tang, S. (2020). We are More than Our Joints: Predicting how 3D Bodies Move. ArXiv, abs/2012.00619.

[14] Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36, 1325-1339.

[15] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M.J. (2015). SMPL: a skinned multi-person linear model. ACM Trans. Graph., 34, 248:1-248:16.

[16] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 5026-5033. IEEE, 2012.