SPATIO-TEMPORAL NORMALIZED JOINT COORDINATES AS FEATURES FOR SKELETON-BASED HUMAN ACTION RECOGNITION

FAKHRUL ANIQ HAKIMI BIN NASRUL ‘ALAM

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Philosophy

Malaysia-Japan International Institute of Technology
Universiti Teknologi Malaysia

MAY 2022

DEDICATION

This thesis is dedicated to my parents.

For their endless love, support, and encouragement.

ACKNOWLEDGEMENT

In the name of Allah, the most gracious and the most merciful. First and foremost, I am thankful to Almighty Allah for giving me the strength, knowledge, ability, and opportunity to undertake this study and complete it satisfactorily.

Secondly, I would like to thank my respected supervisor, Prof. Madya Ts. Dr. Mohd Ibrahim Bin Shapiai, Centre for Artificial Intelligence & Robotics (CAIRO), Malaysia-Japan International Institute of Technology, UTM, whose guidance and professional attitude are appreciable in completing this thesis. Special shout out to Khaleeda, Amirah, Sufian, and all my friends for their encouragement, physically and mentally. Their intakes and advice have also been crucial in guiding me to complete this research.

Last but not least, I want to express my most enormous gratitude to my family, especially my parents, Nasrul 'Alam Bin Nasiruddin and Fadzilah Binti Hashim for their endless support and love. They have raised and taken care of me throughout my life. I am where I am right now because of their utmost support, and they are a big part of my life.

ABSTRACT

Human Action Recognition (HAR) is critical in video monitoring, human-computer interaction, video comprehension, and virtual reality. While significant progress has been made in the HAR domain in recent years, developing an accurate, fast, and efficient system for video action recognition remains a challenge due to a variety of obstacles, such as changes in camera viewpoint, occlusions, background, and motion speed. In general, the action recognition model learns spatial and temporal features in order to classify human actions. The state-of-the-art approaches to deep learning skeleton-based action recognition rely primarily on Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN). RNN-based action recognition methods only model the long-term contextual information in the temporal domain. In return, they neglect the spatial configurations of articulated skeletons where the joints are strongly discriminative. Therefore, it is challenging to extract high-level features. In contrast, action recognition based on CNNs is incapable of modelling long-term temporal dependency. Typically, implementations stack a limited number of frames and convert them into images to represent spatio-temporal information. However, this approach is susceptible to information loss during the conversion process. This study proposes STEM-Coords as pre-processing and features extraction technique, to effectively represent spatio-temporal features using joint coordinates from a human pose. The feature set comprised normalized joint coordinates and their respective speed was represented tabularly as input for the Neural Oblivious Decision Ensemble (NODE) classification model. The proposed STEM-Coords was validated on three benchmark datasets KTH, RealWorld HAR, and MSR DailyActivity 3D. Our method outperformed the state-of-the-art approaches on every dataset with 97.3%, 99.3%, and 97.4% accuracy rates, respectively. The results demonstrated that our proposed method effectively and efficiently represents spatio-temporal information while maintaining robustness to partial occlusion, anthropometrically, and view-invariant..

ABSTRAK

Pengecaman Tindakan Manusia (HAR) adalah penting dalam pemantauan video, interaksi manusia-komputer, pemahaman video dan realiti maya. Walaupun kemajuan ketara telah dicapai di dalam domain HAR dalam beberapa tahun kebelakangan ini, pembangunan sistem yang tepat, pantas dan cekap untuk pengecaman tindakan manusia menggunakan video kekal mencabar disebabkan oleh pelbagai halangan, antaranya termasuk perubahan dalam sudut pandang kamera, halangan pandangan, latar belakang dan kelajuan gerakan. Secara amnya, model pengecaman tindakan mempelajari ciri ruang dan temporal untuk mengklasifikasikan tindakan manusia. Pendekatan tercanggih untuk pengecaman tindakan manusia berasaskan rangka *deep-learning* bergantung terutamanya pada Rangkaian Neural Berulang (RNN) atau Rangkaian Neural Konvolusi (CNN). Kaedah pengecaman tindakan manusia berasaskan RNN hanya memodelkan maklumat kontekstual jangka panjang dalam domain temporal. Oleh itu, ia mengabaikan konfigurasi rangka badan manusia dalam domain ruangan di mana ianya sangat diskriminatif. Sehubungan dengan itu, adalah sangat mencabar untuk mengekstrak ciri-ciri berkualiti tinggi. Sebaliknya, pengecaman tindakan manusia berasaskan CNN tidak mampu memodelkan ciri temporal jangka panjang. Pelaksanaannya adalah berdasarkan penyusunan bilangan bingkai video yang terhad dan penukaran kepada bentuk imej bagi mewakili maklumat ruangan dan temporal. Walau bagaimanapun, pendekatan ini terdedah kepada kehilangan maklumat semasa proses penukaran imej. Kajian ini mencadangkan STEM-Coords sebagai pra-pemprosesan dan teknik pengekstrakan ciri, untuk mewakili ciri ruangan dan temporal dengan berkesan menggunakan koordinat daripada rangka manusia. Set ciri terdiri daripada koordinat sendi ternormal dan kelajuan sebagai data input untuk model klasifikasi *Neural Oblivious Decision Ensemble* (NODE). STEM-Coords yang dicadangkan disahkan pada tiga set data penanda aras KTH, RealWorld HAR dan MSR DailyActivity 3D. Kaedah ini mengatasi pendekatan terkini pada setiap set data dengan kadar ketepatan 97.3%, 99.3% dan 97.4%. Hasil kajian ini menunjukkan bahawa kaedah yang dicadangkan adalah berkesan dan cekap untuk mewakili maklumat ruangan dan temporal sementara juga teguh kepada oklusi separa, antropometrik dan perubahan pandangan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2D	-	Two-Dimension
3D	-	Three-Dimension
4D	-	Four-Dimension
ANN	-	Artificial Neural Network
BOVW	-	Bag-of-Visual-Words
CNN	-	Convolutional Neural Network
DBN	-	Dynamic Bayesian Network
DMM	-	Depth Motion Maps
DNN	-	Deep Neural Network
EM	-	Expectation Maximization
FF	-	Full Feature
FS	-	Feature Selection
GCN	-	Graph Convolutional Network
GMM	-	Gaussian Mixture Model
GPU	-	Graphical Processing Units
HAR	-	Human Action Recognition
HMM	-	Hidden Markov Model
HOF	-	Histogram Of Optical Flow
HOG	-	Histogram Of Oriented Gradients
HOG3D	-	Histogram Of Three-Dimensional Oriented Gradients
HOJ3D	-	Histograms Of Three-Dimensional Joint
HON4D	-	Histogram Of Oriented 4D Normal
LDP	-	Local Depth Pattern
LKT	-	Lucas-Kanade-Tomasi
LOP	-	Local Occupancy Pattern
LSTM	-	Long Short
MEI	-	Motion-Energy Image
MHI	-	Motion-History Image
MLP	-	Multilayer Perceptron
NLP	-	Natural Language Processing

NODE	-	Neural Oblivious Decision Ensemble
P2RN	-	Pose-Guided Recurrent Network
PCA	-	Principal Component Analysis
RF	-	Random Forest
RGB	-	Red, Green, Blue
RGBD	-	Red, Green, Blue, Depth
RNN	-	Recurrent Neural Network
ROI	-	Region Of Interest
SDK	-	Software Development Kit
STIP	-	Space-Time Interest Point
STV	-	Spatiotemporal Volume
SVM	-	Support Vector Machine
VLAD	-	Vector Of Locally Aggregated Descriptor
WHDMM	-	Weighted Hierarchical Depth Motion Maps

LIST OF SYMBOLS

I^k	-	k^{th} frame
J^k	-	instances set of body joints in k^{th} frame
X_i	-	Normalized joint coordinates x
Y_i	-	Normalized joint coordinates y
Δ_i	-	The speed of joint coordinates

CHAPTER 1

INTRODUCTION

1.1 Problem Background

In recent years, Human Action Recognition (HAR) has emerged as a significant area of study in computer vision. It is used in a variety of applications, including human-computer interaction [1], autonomous driving vehicles [2], video surveillance [3], e-health [4], and patient tracking [5].

The primary goal of HAR is to interpret human behavior and actions using sensors or visual data. The HAR process is typically composed of four primary steps: data acquisition, pre-processing, feature extraction, and classification. Data acquisition is the process of obtaining human data from any source input. Pre-processing is the process of eliminating redundant, irrelevant, or noisy features in order to enhance the selected feature set. Two examples of pre-processing techniques are feature normalization and feature selection. Meanwhile, feature extraction is the process of transforming data into processable features while retaining the discriminative information in the original dataset. The last step is classification, which predicts an action class label based on the given data.

There are three main categories of HAR approaches: vision-based action recognition, sensor-based action recognition, and multimodal action recognition [6, 7]. The primary distinction between vision-based and the other two categories is that vision-based approaches utilize 2D or 3D data in the form of images or videos. In contrast, sensor-based methods use time-serial data readings from wearable sensors [7]. Wearable devices such as smartphones, smart watches, and fitness wristbands have been developed in recent years. They are equipped with microprocessors and sensors that enable computation and communication.

Wearable devices have several limitations, the most significant being that they must typically be worn and operated continuously. As a result, specific technical specifications are required, such as battery life, sensor size, and performance [8]. This may pose difficulties in terms of readiness and deployability for real-time applications. Additionally, they may be inefficient or inappropriate for use in specific scenarios, such as crowd applications or others. These limitations, however, do not apply to HAR based on computer vision. Instead, the implementation applies to various applications without complicated technical requirements or constraints. Typically, the vision-based HAR algorithm generates a label after observing the entirety of a human action being performed in a video. In computer vision, the term "human action" refers to various movements ranging from simple joint movement to complex joint movements involving multiple joints and the human body. However, video-based classification has progressed more slowly than expected due to various factors, including the high computational cost. Besides that, the datasets for this application are limited because of the difficulties of collecting, annotating, and storing videos.

Researchers have published numerous studies on action recognition using images or video data since approximately 1980 [9, 10]. They have frequently followed or been inspired by elements of the operating principle of the human vision system. The human vision system receives visual information about an object's movement, shape, and change over time. The observations are passed into the perception system for recognition. Numerous researchers have investigated the biophysical processes underlying the human recognition system in order to develop computer vision systems with comparable performance. However, due to various constraints, such as environmental complexity, scale variation, non-rigid shapes, background clutter, viewpoint variation, and occlusions, computer vision systems are unable to fully realize some fundamental aspects of the human vision system.

1.2 Problem Statement

Although significant progress has been made in the HAR domain in recent years, developing an accurate, fast, and efficient system for video action recognition remains challenging due to various obstacles, including changes in camera viewpoint, occlusions, background, and motion speed. Historically, video-based action recognition techniques have emphasized the extraction of handcrafted global features like silhouette, shapes, and optical flow [11-23]. However, due to its sensitivity to noise, occlusions, and viewpoint changes, it has become increasingly obsolete. Moreover, silhouettes and shapes are now more uncomplicated to obtain without sophisticated algorithms due to the advancement of the modern RGBD camera.

Therefore, research has shifted their attention to handcrafted local features to resolve the issues caused by global features. It has been demonstrated that most local features are robust to noise and partial occlusions. Numerous local representations for action recognition, including spatio-temporal interest points (STIP) [24-29] and Dense trajectories [30-33], have been proposed and successfully implemented. However, although these local features produce excellent results in HAR, they come with several limitations. One of the limitations is the lack of stable discriminative interest points because it is difficult to identify and maintain the stability of interest points with the number of points discovered. As a result, these techniques remain limited to minor point detection or low-resolution video.

To overcome the challenges faced by global and local features, researchers try to take advantage of the development of low-cost depth sensors [34]. Previously, studies utilizing depth sensors were limited due to their high cost and technical complexity. Depth sensors generate precise depth maps of human action. Furthermore, most depth sensors incorporate real-time skeleton estimation and tracking algorithms, which simplifies the collection of skeleton information. This is a high-level representation of the human body appropriate for the motion analysis problem. Thus, utilizing depth maps and skeletal information can overcome the limitations of conventional RGB-based approaches. As a result, numerous depth sensor approaches have been proposed [35-37]. However, as standalone features, depth maps are ineffective at recognizing human actions. Due to the absence of temporal information, it is difficult to distinguish between dynamic actions such as running, walking, and jogging. As a result, depth maps are frequently combined with other features such as skeletal information or handcrafted RGB video features. Additionally, depth sensors have some significant limitations. For example, low-cost depth sensors cannot operate in direct sunlight and have a limited range and field of vision. As a result, the data extracted from depth sensors are extremely noisy, necessitating additional pre-processing.

To address this issue, researchers developed a pose estimation network that can generate skeleton information directly from videos. Skeleton data derived from the pose estimation network can capture the motions of human skeleton joints and are illumination invariant [38]. However, skeleton data require pre-processing because they are not view-invariant and are susceptible to anthropometric variability. As a result, the features have lower discriminatory power. Several handcrafted pose estimation approaches have used more sophisticated geometric tools to model human actions [39, 40]. Because these descriptors are derived using invariant features such as the distance between joints, angles, and transformation matrices, they are implicitly unaffected by viewpoint variability. Alternatively, applying an alignment pre-processing step can achieve similar results before performing the descriptor computation, reducing the system's overall complexity.

While these representations have demonstrated their efficacy in terms of computation time and accuracy, it has been demonstrated that handcrafted features perform well on a limited number of datasets [41]. For example, handcrafted features are optimized for a specific dataset and may not be applicable to other datasets. This makes it difficult for action recognition to be generalized into broader applications. Additionally, because handcrafted methods are effective at avoiding overfitting, they may be unable to learn from larger datasets. However, with the increased availability of large benchmark datasets in recent years, the future research trend is more likely to shift toward using deep learning features.

Numerous deep learning approaches have been proposed for recognizing human actions using skeletons. The most frequently used deep learning architectures are CNN and RNN. However, few studies investigate the use of alternative network architectures. Temporal information can be extracted from spatial sequences using RNN architectures. A significant disadvantage of their approach is the exploding and vanishing gradient problem and the difficulty of parallelizing their training.

Therefore, a more advanced RNN, the LSTM, is used to enable training on long sequences. However, even if LSTM networks are designed to explore long-term dependencies, it is still challenging to learn the information in an entire sequence with numerous timestamps [42, 43]. These RNN-based action recognition methods only model the long-term contextual information in the temporal domain. In return, they neglect the spatial configurations of articulated skeletons where the joints are strongly discriminative. Therefore, it is difficult for LSTM networks to extract high-level features [44, 45].

On the other hand, Convolutional Neural Networks (CNNs) have demonstrated tremendous potential for image pattern recognition [46]. However, for video action recognition, it still lacks the capacity to model the long-term temporal dependency of the entire video [47]. Therefore, the implementations typically focus on optimizing spatial feature extraction through various normalization methods. Some approaches make use of spatio-temporal characteristics. However, the extraction method involves a highly complex combination of spatial and temporal features. The implementation is frequently based on the conversion of skeleton sequences to images in which the spatio-temporal information is reflected in the image properties, such as color and texture [48]. One disadvantage of the approach is that it is unavoidable for temporal information to be lost during the data conversion.

1.3 Research Goal

In accordance with the stated problem statement, the primary goal of this study is to develop a deep learning skeleton-based approach for an action recognition system capable of accurately predicting actions from video sequences by efficiently and effectively representing spatio-temporal features using joint coordinates from a human pose that are robust to part occlusion, and anthropometric-, illumination-, view-invariant.

1.3.1 Research Objectives

The objectives of the research are:

(a) To develop a skeleton-based action recognition model by combining a Residual Network (ResNet) pose estimation model with a Neural Oblivious Decision Ensemble (NODE) architecture as the classification network.

(b) To develop pre-processing and feature extraction techniques for skeleton joint location in order to enable temporal and spatial modeling in the feature set represented tabularly for the classification model (a).

(c) To validate the effectiveness of proposed method (b) by conducting performance analysis of the classification network in (a) in terms of overall and per class classification over three benchmark datasets: KTH, RealWorld HAR, and MSR DailyActivity 3D.

1.4 Research Scope

Human action interpretation from a video is a hot research topic these days. The research conducted in this domain can be classified into two subfields: action recognition and detection. After processing the video, an action label is assigned to it in action recognition. Meanwhile, action detection identifies and locates the action within the video frame. This research focuses on detecting and recognizing human actions based on deep learning architecture. Meanwhile, humans are detected and localized in the frame in spatial and temporal domains by using a pose estimation network that extracts skeletal information for the prediction of action labels.

Most action classification algorithms can be classified into three types: template-based, generative, or discriminative models. The term "template-based" refers to a technique for identifying the shared characteristics of a specific action. This characteristic may consist of two-dimensional or three-dimensional images, volumes, or a sequence of view models. The generative model is a technique for determining the most likely label prediction by calculating the joint probabilities of input X and class labels Y using the Bayes rule. On the other hand, the discriminative model can directly determine the label for prediction by utilizing advanced machine learning algorithms. RNN and CNN are the two most frequently used discriminative models in the literature. This study uses the deep learning Neural Oblivious Decision Ensemble (NODE) architecture to develop our classification model.

There are two types of input modalities: vision-based and sensor-based. Sensor-based classification refers to the process of classifying actions using data from inertial sensors such as an accelerometer and a gyroscope. Although it is rich in motion data, it lacks spatial information. Therefore, we concentrate on utilizing RGB videos as our input in this study. RGB video contains a wealth of spatiotemporal information critical for recognizing human action, particularly in dynamic and static activities. We use a pose estimation network to extract the discriminative spatial configuration of articulated skeletons. Additionally, we can model the temporal dependencies by considering skeletons in multiple sequences of frames.

Human actions can be classified into four broad categories based on their context. The first category is "gestures," which denotes a precise movement of a body part, such as "raising a leg." The second category is "action," which refers to a collection of a person’s coordinated gestures, such as "walking" or "waving." The third category is "Interaction," which encompasses situations involving two or more people, objects, or both simultaneously. For example, pushing another person is a two-person interaction, whereas lifting a box is a human-object interaction. The final category is

"Group Activity," which includes activities multiple individuals participate in, such as a group of people running. This study focuses on recognizing single-person actions, particularly on “action” and “human-object interaction” categories.

1.5 Research Contributions

In general, this research makes two significant contributions:

First, we developed a skeleton-based action recognition model that utilizes spatio-temporal joint coordinates as features. We introduced STEM-Coords, a method of extracting spatial and temporal joint coordinates information from a set of window-frames. This method includes eliminating redundant joints and normalizing the remaining joints, which we refer to as "active joints," thereby enhancing the feature saliency. Therefore, we conducted an extensive analysis to demonstrate the effectiveness of STEM-Coords. We utilize a simple and robust SimpleBaseline pose estimation network to obtain raw skeletal data. Due to the lightweight of the classification model, real-time HAR implementation is possible.

Second, our classification model is based on the Neural Oblivious Decision Ensemble (NODE) architecture. It is a recent state-of-the-art deep learning model for tabular data. Our classification system is the first architecture implementation in the literature for any application. The developed model achieves state-of-the-art performance on three challenging benchmarks: KTH, RealWorld HAR, and MSR DailyActivity 3D. We conducted a comprehensive performance analysis of the classification model against various state-of-the-art approaches. This analysis demonstrated the effectiveness of the model for individual action classes and overall actions.

1.6 Thesis Organization

The remainder of this thesis follows the following structure: Chapter 2 conducts a comprehensive review of the literature in the field of human action recognition. This section discusses HAR technology and research evolution, from conventional to contemporary approaches. First, it discusses the three broad categories of HAR: template-based, generative, and discriminative models. The discussion then narrows to the discriminative model and discusses the various input modalities used in the literature, including wearable sensor-based and vision-based. Next, this chapter discusses the feature representation for the vision-based action recognition system. Finally, this chapter concludes with a critical review of the relevant literature.

Chapter 3 describes the detailed methodology of our action recognition system. It begins by providing an overview of the overall model. The overview consists of several blocks that represent their primary function. The function of the blocks is discussed in detail by sections: pose estimation, feature pre-processing and extraction, and classification. The latter part of this chapter discusses the experimental procedures and parameters.

Chapter 4 provides an in-depth analysis of the effectiveness of our feature extraction method, STEM-Coords, and the performance evaluation of the classification model. The analysis is based on three experiments: 1) Investigating the effect of removing redundant joints, 2) Investigating the effect of incorporating temporal information, and 3) Investigating the performance of the model in comparison to other state-of-the-art approaches. Three challenging datasets were used for the experiments: KTH, RealWorld HAR, and MSR DailyActivity3D. Finally, chapter 5 summarizes the research by providing conclusions and outlining the recommendation and future direction of the research.

REFERENCES

1. Poppe, R., *A survey on vision-based human action recognition.* Image and vision computing, 2010. **28**(6): p. 976-990.
2. Chen, L., et al., *Survey of pedestrian action recognition techniques for autonomous driving.* Tsinghua Science and Technology, 2020. **25**(4): p. 458-470.
3. Jin, C.-B., et al. *Real-time human action recognition using CNN over temporal images for static video surveillance cameras*. in *Pacific Rim Conference on Multimedia*. 2015. Springer.
4. Bao, J., M. Ye, and Y. Dou. *Mobile phone-based internet of things human action recognition for E-health*. in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. 2016. IEEE.
5. Chou, E., et al., *Privacy-preserving action recognition for smart hospitals using low-resolution depth images.* arXiv preprint arXiv:1811.09950, 2018.
6. Yurur, O., C.H. Liu, and W. Moreno, *A survey of context-aware middleware designs for human activity recognition.* IEEE Communications Magazine, 2014. **52**(6): p. 24-31.
7. Ranasinghe, S., F. Al Machot, and H.C. Mayr, *A review on applications of activity recognition systems with regard to performance and evaluation.* International Journal of Distributed Sensor Networks, 2016. **12**(8): p. 1550147716665520.
8. Chen, L., et al., *Sensor-based activity recognition.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012. **42**(6): p. 790-808.
9. Marr, D. and L. Vaina, *Representation and recognition of the movements of shapes.* Proceedings of the Royal Society of London. Series B. Biological Sciences, 1982. **214**(1197): p. 501-524.
10. Hester, C.F. and D. Casasent, *Multivariant technique for multiclass pattern recognition.* Applied Optics, 1980. **19**(11): p. 1758-1761.

11. Blank, M., et al. *Actions as space-time shapes*. in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. 2005. IEEE.
12. Wren, C.R., et al., *Pfinder: Real-time tracking of the human body.* IEEE Transactions on pattern analysis and machine intelligence, 1997. **19**(7): p. 780-785.
13. Koller, D., et al. *Towards robust automatic traffic scene analysis in real-time*. in *Proceedings of 12th International Conference on Pattern Recognition*. 1994. IEEE.
14. Stauffer, C. and W.E.L. Grimson. *Adaptive background mixture models for real-time tracking*. in *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*. 1999. IEEE.
15. Veeraraghavan, A., A.R. Chowdhury, and R. Chellappa. *Role of shape and kinematics in human movement analysis*. in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* 2004. IEEE.
16. Veeraraghavan, A., R. Chellappa, and A.K. Roy-Chowdhury. *The function space of an activity*. in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. 2006. IEEE.
17. Bobick, A. and J. Davis. *An appearance-based representation of action*. in *Proceedings of 13th International Conference on Pattern Recognition*. 1996. IEEE.
18. Bobick, A.F. and J.W. Davis, *The recognition of human movement using temporal templates.* IEEE Transactions on pattern analysis and machine intelligence, 2001. **23**(3): p. 257-267.
19. Ikizler, N. and P. Duygulu. *Human action recognition using distribution of oriented rectangular patches*. in *Workshop on Human Motion*. 2007. Springer.
20. Lucas, B.D. and T. Kanade. *An iterative image registration technique with an application to stereo vision*. 1981. Vancouver, British Columbia.
21. Shi, J. *Good features to track*. in *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. 1994. IEEE.
22. Lu, X., Q. Liu, and S. Oe. *Recognizing non-rigid human actions using joints tracking in space-time*. in *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.* 2004. IEEE.

23. Efros, A.A., et al. *Recognizing action at a distance*. in *Computer Vision, IEEE International Conference on*. 2003. IEEE Computer Society.

24. Peng, X., et al., *Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice.* Computer Vision and Image Understanding, 2016. **150**: p. 109-125.

25. Harris, C. and M. Stephens. *A combined corner and edge detector*. in *Alvey vision conference*. 1988. Citeseer.

26. Laptev, I., *On space-time interest points.* International journal of computer vision, 2005. **64**(2): p. 107-123.

27. Kadir, T. and M. Brady, *Saliency, scale and image description.* International Journal of Computer Vision, 2001. **45**(2): p. 83-105.

28. Oikonomopoulos, A., I. Patras, and M. Pantic, *Spatiotemporal salient points for visual recognition of human actions.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2006. **36**(3): p. 710-719.

29. Dollár, P., et al. *Behavior recognition via sparse spatio-temporal features*. in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. 2005. IEEE.

30. Tran, D. and A. Sorokin. *Human activity recognition with metric learning*. in *European conference on computer vision*. 2008. Springer.

31. Ke, Y., R. Sukthankar, and M. Hebert. *Efficient visual event detection using volumetric features*. in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. 2005. IEEE.

32. Dalal, N. and B. Triggs. *Histograms of oriented gradients for human detection*. in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. 2005. Ieee.

33. Scovanner, P., S. Ali, and M. Shah. *A 3-dimensional sift descriptor and its application to action recognition*. in *Proceedings of the 15th ACM international conference on Multimedia*. 2007.

34. Shotton, J., et al. *Real-time human pose recognition in parts from single depth images*. in *CVPR 2011*. 2011. Ieee.

35. Oreifej, O. and Z. Liu. *Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013.

36. Jalal, A., S. Kamal, and D. Kim, *A Depth Video-based Human Detection and Activity Recognition using Multi-features and Embedded Hidden Markov Models for Health Care Monitoring Systems.* International Journal of Interactive Multimedia & Artificial Intelligence, 2017. **4**(4).
37. Luo, J., W. Wang, and H. Qi. *Group sparsity and geometry constrained dictionary learning for action recognition from depth maps*. in *Proceedings of the IEEE international conference on computer vision*. 2013.
38. Han, F., et al., *Space-time representation of people based on 3D skeletal data: A review.* Computer Vision and Image Understanding, 2017. **158**: p. 85-105.
39. Vemulapalli, R., F. Arrate, and R. Chellappa. *Human action recognition by representing 3d skeletons as points in a lie group*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.
40. Evangelidis, G., G. Singh, and R. Horaud. *Skeletal quads: Human action recognition using joint quadruples*. in *2014 22nd International Conference on Pattern Recognition*. 2014. IEEE.
41. Wang, L., D.Q. Huynh, and P. Koniusz, *A comparative review of recent kinect-based action recognition algorithms.* IEEE Transactions on Image Processing, 2019. **29**: p. 15-28.
42. Gu, J., G. Wang, and T. Chen, *Recurrent highway networks with language cnn for image captioning.* arXiv preprint arXiv:1612.07086, 2016.
43. Bordes, A., et al., *Large-scale simple question answering with memory networks.* arXiv preprint arXiv:1506.02075, 2015.
44. Pascanu, R., et al., *How to construct deep recurrent neural networks.* arXiv preprint arXiv:1312.6026, 2013.
45. Sainath, T.N., et al. *Convolutional, long short-term memory, fully connected deep neural networks*. in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2015. IEEE.
46. Taigman, Y., et al. *Closing the gap to human-level performance in face verification. deepface*. in *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR).*
47. Wang, L., et al. *Temporal segment networks: Towards good practices for deep action recognition*. in *European conference on computer vision*. 2016. Springer.

48. Ke, Q., et al. *A new representation of skeleton sequences for 3d action recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
49. Zhang, S., et al., *A review on human activity recognition using vision-based method.* Journal of healthcare engineering, 2017. **2017**.
50. Almaslukh, B., A.M. Artoli, and J. Al-Muhtadi, *A robust deep learning approach for position-independent smartphone-based human activity recognition.* Sensors, 2018. **18**(11): p. 3726.
51. Aljarrah, A.A. and A.H. Ali, *Human Activity Recognition by Deep Convolution Neural Networks and Principal Component Analysis*, in *Further Advances in Internet of Things in Biomedical and Cyber Physical Systems*. 2021, Springer. p. 111-133.
52. Ordóñez, F.J. and D. Roggen, *Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition.* Sensors, 2016. **16**(1): p. 115.
53. Kasnesis, P., C.Z. Patrikakis, and I.S. Venieris. *PerceptionNet: a deep convolutional neural network for late sensor fusion*. in *Proceedings of SAI Intelligent Systems Conference*. 2018. Springer.
54. Damirchi, H., R. Khorrambakht, and H.D. Taghirad. *ARC-net: Activity recognition through capsules*. in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2020. IEEE.
55. Klaser, A., M. Marszałek, and C. Schmid. *A spatio-temporal descriptor based on 3d-gradients*. in *BMVC 2008-19th British Machine Vision Conference*. 2008. British Machine Vision Association.
56. Willems, G., T. Tuytelaars, and L. Van Gool. *An efficient dense and scale-invariant spatio-temporal interest point detector*. in *European conference on computer vision*. 2008. Springer.
57. Wang, H., et al. *Action recognition by dense trajectories. Computer Vision and Pattern Recognition (CVPR)*. in *2011 IEEE Conference on*. 2011.
58. Farnebäck, G. *Two-frame motion estimation based on polynomial expansion*. in *Scandinavian conference on Image analysis*. 2003. Springer.
59. Li, W., Z. Zhang, and Z. Liu. *Action recognition based on a bag of 3d points*. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. 2010. IEEE.

60. Yang, X., C. Zhang, and Y. Tian. *Recognizing actions using depth motion maps-based histograms of oriented gradients*. in *Proceedings of the 20th ACM international conference on Multimedia*. 2012.
61. Shotton, J., et al., *Real-time human pose recognition in parts from single depth images.* Communications of the ACM, 2013. **56**(1): p. 116-124.
62. Chen, H., et al., *A novel hierarchical framework for human action recognition.* Pattern Recognition, 2016. **55**: p. 148-159.
63. Wang, J., et al., *Learning actionlet ensemble for 3D human action recognition.* IEEE transactions on pattern analysis and machine intelligence, 2013. **36**(5): p. 914-927.
64. Shahroudy, A., et al., *Multimodal multipart learning for action recognition in depth videos.* IEEE transactions on pattern analysis and machine intelligence, 2015. **38**(10): p. 2123-2129.
65. Shahroudy, A., et al. *Ntu rgb+ d: A large scale dataset for 3d human activity analysis*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
66. Liu, J., et al., *An ultra-fast human detection method for color-depth camera.* Journal of Visual Communication and Image Representation, 2015. **31**: p. 177-185.
67. Liu, J., et al., *Detecting and tracking people in real time with RGB-D camera.* Pattern Recognition Letters, 2015. **53**: p. 16-23.
68. Liu, J., et al. *Real-time human detection and tracking in complex environments using single RGBD camera*. in *2013 IEEE International Conference on Image Processing*. 2013. ieee.
69. Rabiner, L. and B. Juang, *An introduction to hidden Markov models.* ieee assp magazine, 1986. **3**(1): p. 4-16.
70. Freund, Y. and R.E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting.* Journal of computer and system sciences, 1997. **55**(1): p. 119-139.
71. Xia, L., C.-C. Chen, and J.K. Aggarwal. *View invariant human action recognition using histograms of 3d joints*. in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. 2012. IEEE.

72. Wang, J., et al. *Mining actionlet ensemble for action recognition with depth cameras*. in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012. IEEE.

73. Yang, X. and Y.L. Tian. *Eigenjoints-based action recognition using naive-bayes-nearest-neighbor*. in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. 2012. IEEE.

74. Zhang, C. and Y. Tian, *RGB-D camera-based daily living activity recognition.* Journal of computer vision and image processing, 2012. **2**(4): p. 12.

75. Sempena, S., N.U. Maulidevi, and P.R. Aryan. *Human action recognition using dynamic time warping*. in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*. 2011. IEEE.

76. Ghorbel, E., et al., *Kinematic Spline Curves: A temporal invariant descriptor for fast action recognition.* Image and Vision Computing, 2018. **77**: p. 60-71.

77. Bengio, Y., P. Simard, and P. Frasconi, *Learning long-term dependencies with gradient descent is difficult.* IEEE transactions on neural networks, 1994. **5**(2): p. 157-166.

78. Pascanu, R., T. Mikolov, and Y. Bengio. *On the difficulty of training recurrent neural networks*. in *International conference on machine learning*. 2013. PMLR.

79. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.

80. Cho, K., et al., *On the properties of neural machine translation: Encoder-decoder approaches.* arXiv preprint arXiv:1409.1259, 2014.

81. Du, Y., W. Wang, and L. Wang. *Hierarchical recurrent neural network for skeleton based action recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

82. Du, Y., Y. Fu, and L. Wang, *Representation learning of temporal dynamics for skeleton-based action recognition.* IEEE Transactions on Image Processing, 2016. **25**(7): p. 3010-3022.

83. Veeriah, V., N. Zhuang, and G.-J. Qi. *Differential recurrent neural networks for action recognition*. in *Proceedings of the IEEE international conference on computer vision*. 2015.

84. Zhu, W., et al. *Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks*. in *Proceedings of the AAAI conference on artificial intelligence*. 2016.
85. Liu, J., et al. *Spatio-temporal lstm with trust gates for 3d human action recognition*. in *European conference on computer vision*. 2016. Springer.
86. Song, S., et al. *An end-to-end spatio-temporal attention model for human action recognition from skeleton data*. in *Proceedings of the AAAI conference on artificial intelligence*. 2017.
87. Bai, S., J.Z. Kolter, and V. Koltun, *An empirical evaluation of generic convolutional and recurrent networks for sequence modeling.* arXiv preprint arXiv:1803.01271, 2018.
88. Zhang, Y., et al., *Towards end-to-end speech recognition with deep convolutional neural networks.* arXiv preprint arXiv:1701.02720, 2017.
89. Gehring, J., et al., *A convolutional encoder model for neural machine translation.* arXiv preprint arXiv:1611.02344, 2016.
90. Adel, H. and H. Schütze, *Exploring different dimensions of attention for uncertainty detection.* arXiv preprint arXiv:1612.06549, 2016.
91. Wang, P., et al., *Action recognition based on joint trajectory maps with convolutional neural networks.* Knowledge-Based Systems, 2018. **158**: p. 43-53.
92. Li, C., et al., *Joint distance maps based action recognition with convolutional neural networks.* IEEE Signal Processing Letters, 2017. **24**(5): p. 624-628.
93. Liu, M., H. Liu, and C. Chen, *Enhanced skeleton visualization for view invariant human action recognition.* Pattern Recognition, 2017. **68**: p. 346-362.
94. Yan, S., Y. Xiong, and D. Lin. *Spatial temporal graph convolutional networks for skeleton-based action recognition*. in *Thirty-second AAAI conference on artificial intelligence*. 2018.
95. Shi, L., et al. *Two-stream adaptive graph convolutional networks for skeleton-based action recognition*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
96. Li, M., et al. *Actional-structural graph convolutional networks for skeleton-based action recognition*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

97. Laptev, I. and B. Caputo, *Recognizing human actions: a local svm approach. In null.* IEEE, 2004. **5**: p. 32-36.

98. Sztyler, T., *Sensor-based human activity recognition: Overcoming issues in a real world setting*. 2019: Universitaet Mannheim (Germany).

99. Ellis, C., et al., *Exploring the trade-off between accuracy and observational latency in action recognition.* International Journal of Computer Vision, 2013. **101**(3): p. 420-436.

100. Ofli, F., et al. *Berkeley mhad: A comprehensive multimodal human action database*. in *2013 IEEE workshop on applications of computer vision (WACV)*. 2013. IEEE.

101. Ke, Q., et al. *Human interaction prediction using deep temporal features*. in *European Conference on Computer Vision*. 2016. Springer.

102. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.

103. Xiao, B., H. Wu, and Y. Wei. *Simple baselines for human pose estimation and tracking*. in *Proceedings of the European conference on computer vision (ECCV)*. 2018.

104. Popov, S., S. Morozov, and A. Babenko, *Neural oblivious decision ensembles for deep learning on tabular data.* arXiv preprint arXiv:1909.06312, 2019.

105. Poggio, T., et al., *Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review.* International Journal of Automation and Computing, 2017. **14**(5): p. 503-519.

106. contributors, W. *Confusion matrix*. 19 May 2022 [cited 2022 20 May]; Available from: https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=1088688630.

107. Hadfield, S. and R. Bowden. *Hollywood 3D: Recognizing actions in 3D natural scenes*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

108. Arunnehru, J., G. Chamundeeswari, and S.P. Bharathi, *Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos.* Procedia computer science, 2018. **133**: p. 471-477.

109. Jaouedi, N., N. Boujnah, and M.S. Bouhlel, *A new hybrid deep learning model for human action recognition.* Journal of King Saud University-Computer and Information Sciences, 2020. **32**(4): p. 447-453.

110. Srimath, S., et al. *Human Activity Recognition from RGB Video Streams Using 1D-CNNs*. in *2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI)*. 2021. IEEE.

111. Snoun, A., et al., *Towards a deep human activity recognition approach based on video to image transformation with skeleton data.* Multimedia Tools and Applications, 2021. **80**(19): p. 29675-29698.

112. Shahroudy, A., et al., *Deep multimodal feature analysis for action recognition in rgb+ d videos.* IEEE transactions on pattern analysis and machine intelligence, 2017. **40**(5): p. 1045-1058.

113. Zhu, J., et al., *Action machine: Rethinking action recognition in trimmed videos.* arXiv preprint arXiv:1812.05770, 2018.

114. Zhu, Y., W. Chen, and G. Guo, *Fusing multiple features for depth-based action recognition.* ACM Transactions on Intelligent Systems and Technology (TIST), 2015. **6**(2): p. 1-20.

LIST OF PUBLICATIONS

(a) **Nasrul 'Alam, F.A.H., et al**., Skeleton-Based Action Recognition with Joint Coordinates as Feature Using Neural Oblivious Decision Ensembles, in New Trends in Intelligent Software Methodologies, Tools and Techniques. 2021