Dynamic 3D Hand Gesture Recognition

Soft Computing (IT402)

Submitted in partial fulfilment of the requirements for the degree of

Bachelor of Technology

In

Information Technology

by

Utkarsh Meshram (181IT250)

Bhagyashri Bhamare (181IT111)

Chinmayi C. Ramakrishna (181IT113)

Department of Information Technology

National Institute of Technology Karnataka

Surathkal, Mangalore -575025

April, 2021

Acknowledgement

History of all great work as a witness that no great work is ever done with the help and support of one’s surrounding and close quarters. We would like to express our sincere gratitude to our course instructor Mrs. Nagamma Patil to provide us the opportunity to take up this project. We would like to extend our gratitude to our project mentor Mr. Pandian C for giving us suggestions on improvement.

Also, a special thanks to the authors of the research paper, “Dynamic 3D Hand Gesture Recognition by Learning Weighted Depth Motion Maps”. This research paper gave us a direction to implement the project and helped us gain necessary knowledge.

Declaration

We hereby *declare* that the *Soft Computing (IT402) Report* entitled Dynamic 3d Hand Gesture Hand Recognition which is being submitted to the National Institute of Technology Karnataka Surathkal, in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in the department of Information Technology, is a ***bonafide report of the work carried out by us***. The material contained in this seminar report has not been submitted to any University or Institution for the award of any degree.

Bhagyashri Bhamare (181IT111)
Chinmayi C. Ramakrishna (181IT113)
Utkarsh Meshram (181IT250)

Place: NITK, Surathkal

Date: 14.04.2021

Table of Contents

Topic Name **Page Number**

List of Figures

List of Tables

Tables **Page No.**

Abstract

The past few years have seen the emergence of 3d hand gesture recognition. There is a lot of research done in this field for accurate recognition. Many input devices and sensors have been proposed to detect human gesture and convert it into computer information. Many of these have complex algorithms or consume large computational power or lack good recognition algorithms. Also, hand gesture recognition problems include high intra class variability and low inter class variability. In this paper, a multi level temporal sampling (MTS) has been used. It is based on the motion energy of key frames in depth sequences. As a result, MTS lowers intra class variability and increases inter class variability. The weighted depth motion map (WDMM) is then proposed to extract the spatio-temporal information from generated summarized sequences. For this purpose, an accumulated weighted absolute difference of consecutive frames is used. Additionally, descriptors are used to extract features from WDMM. VLAD Encoding is used to compute the final feature descriptor of the video. The SLFN with ELM method is used for classification. The mentioned methodology is evaluated on four RGB-D datasets of MSR Gesture 3D, SKIG, and MSR Action 3D and NTU.

1. Introduction

Multimodal gestures like hand, face, head, finger and body gestures are used to control the application. Gesture recognition systems can be applied in sign language for hearing impaired people[1], distance learning/tele-teaching assistance, video surveillance and monitoring[2], remote control[3], human-environment interaction[4] and guiding the robots[5]. Hand Gesture Recognition can be decomposed into three main steps of hand detection, feature extraction, and classification. There exist various challenges to achieve accurate results. This is due to complex topology of the hand skeleton and high similarity among fingers. During feature extraction, the model has to be designed to deal with various camera viewpoints, different speed of hand gestures, variable hand sizes, style etc. All these factors can contribute to various shapes relating to the same pose. Feix et al.[6] found 17 different hand shapes that humans commonly use in everyday tasks to perform grasping. Multilevel Temporal Sampling is used to augment input data to recognise gestures of different execution rates. A special type of Depth Motion Maps (DMM) called Weighted Depth Motion Maps (WDMM) is used to compute one single image from a sequence of depth frames. Motion energy of the projected depth maps is accumulated into three projective views. Fabio Dominio et al.[7] analysed different features in a hand gesture and summarised methods for 7 features like distance features, elevation features, palm area features, curvature features etc. In this paper, the HOG and LBP descriptors are employed to extract features from WDMM. HOG computes occurrences of gradient orientation in localised portions of an image. LBP is a texture descriptor that describes the local texture patterns of an image by labelling all image pixels with a binary code. The vector of locally aggregated descriptors VLAD encoding process[8] is incorporated to transform the local features (extracted by HOG and LBP) into a fixed-size vector. Dimensionality reduction of feature vectors is applied with the principal component analysis (PCA) on the VLAD encoded descriptor. The last step was the use of a single hidden layer feed-forward neural network (SLFN) with an extreme learning machine (ELM)[9] for classification of hand gestures.

2. Literature Survey

2.1. Related Work

Recent research has suggested a range of solutions to problems, including the use of accurate instruments to capture 3D hand movements and motion, as well as color gloves with attached sensors to capture real-time hand measurements[10,11]. Their calibration configuration process, on the other hand, is time-consuming and expensive. Shotton et al. [12] proposed the "body skeleton" in 2013 as a way to reliably predict the 3D locations of 20 body joints from depth photographs. Traditional methods for encoding the statuses of hand motion and hand form created feature descriptors in the spatial and temporal dimensions [13,14].Methods based on deep learning are currently considered solutions for efficiently and accurately recognizing and classifying images. Dynamic gesture recognition, in particular, employs deep learning techniques such as [15,16], but they are constrained in terms of real-time execution. The key deep learning-based methods were divided into four classes by Asadi et al.[13] 2D models, motion-based input features, 3D models, and temporal methods. A 2D convolutional neural network (CNN) is used in the first category to extract spatial features from one or more sampled frames of the entire film. Score averaging of the outcomes of sampled frames is used to determine the gesture's name. Finally, the fourth category processes input sequences using temporal processing techniques (such as recurrent neural networks (RNN) with LSTM). For understanding hand movements, Molchanov et al. [14] used both 3D CNN and RNN. Short clips of the entire video were fed to the 3D CNN, and the outputs of the 3D CNN were used as the input to the RNN in their process. In the context of action recognition, Simonyan and Zisserman[17] propose separate CNNs for the spatial and temporal streams that are late-fused and that explicitly use optical flow. A 3D-CNN is used by Tran et al.[18] to evaluate a sequence of short video clips and average the network's responses across all clips.

2.2. Motivation

The consumer has been required to adhere to the machine's interface since the invention of the device. With recent developments in Human Computer Intelligent Interaction (HCII), creating interfaces that mimic human communication has become more feasible. While it is still impossible to construct a universal interface that can manage all modes of human communication, a limited multimodal subset can be created. The aim of gesture recognition researchers is to create a system that can recognize gestures, which are commonly used for communicating information or controlling devices. Camera-based gesture recognition solutions have been widely used in a variety of applications and have the ability to communicate through Human Computer Interaction. The majority of modern interfaces rely solely on one mode of interaction, such as speech or mouse, and seldom attempt to use multiple modes. With the possible exception of the speech recognizer, many of the above interfaces have been effective for a professional user, yet they are usually inefficient as a human-centric type of communication. For example, rather than staring at a terminal full of characters when typing a message, we would rather see the individual and converse with him or her while interacting with a friend. A custom-designed test bed would be created if a researcher decided to investigate various applications and capabilities of a multimodal interface. It's time for a more realistic and adaptable test bed that allows for rapid prototyping.

2.3. Problem Statement

The proposed method's input videos are sequences of hand depth maps. The method's output is the label for each input video. The subject makes one significant hand gesture in each study. The hand region is first standardized to a predetermined scale (to cope with different hand sizes). The hand region is first standardized to a predetermined scale (to cope with different hand sizes). The MST creates three levels of videos based on the main frame extraction method, each with a different fixed number of frames from the original. To create 2D projected images, each depth frame is projected onto three Cartesian planes. Every sequence is divided into shorter clips to account for the temporal detail. To compute one image from a sequence of depth frames, a temporal weighted version of DMM (WDMM) is proposed. Each WDMM is broken down into patches to describe it. The HOG and LBP descriptors are employed to extract features from each patch of $WDMM^{l;v}_{c,p}$, which is the p^{th} patch of the WDMM computed from the c^{th} clip with the temporal level of l in view v. All of the features extracted from one sample are encoded by the VLAD encoding. The SLNF with ELM method is exploited for the classification.

2.4. Objectives

1. Dynamic 3D Human Hand Gesture Recognition on RGB-D videos
2. Use of State of the Art results on public data sets.
3. This Method Learns Human Actions with Aggregating of Spatio-Temporal Description from different representation
4. To achieve the state-of-the-art results on MSR Action 3D datasets and outperforming deep learning results.

3. Proposed Methodology

4. Result and Analysis

5. Conclusion and Future Work

References

[1] Ling Shao, Ziyun Cai, Li Liu, and Ke Lu. Performance evaluation of deep feature learning for rgb-d image/video classification. Information Sciences, 385:266–283, 2017.

[2] Hang Zhou and Qiuqi Ruan. A real-time gesture recognition algorithm on video surveillance. In Signal Processing, 2006 8th International Conference on, volume 3. IEEE, 2006.

[3] Utpal V Solanki and Nilesh H Desai. Hand gesture based remote control for home appliances: Handmote. In Information and Communication Technologies, 2011 World Congress on, pages 419–423. IEEE, 2011.

[4] Siddharth S Rautaray and Anupam Agrawal. Real time gesture recognition system for interaction in dynamic environment. Procedia Technology, 4:595–599, 2012.

[5] Coronado, Enrique & Villalobos, Jessica & Bruno, Barbara & Mastrogiovanni, Fulvio. (2017). Gesture-based Robot Control: Design Challenges and Evaluation with Humans. 10.1109/ICRA.2017.7989321.

[6] Feix, T.; Pawlik, R.; Schmiedmayer, H.B.; Romero, J.; Kragi, D. A comprehensive grasp taxonomy. In Proceedings of Robotics, Science and Systems Conference: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation, Seattle, WA, USA, 28 June–1 July 2009; pp. 2–3.

[7] Dominio F., Marin G., Piazza M., Zanuttigh P. (2014) Feature Descriptors for Depth-Based Hand Gesture Recognition. In: Shao L., Han J., Kohli P., Zhang Z. (eds) Computer Vision and Machine Learning with RGB-D Sensors. Advances in Computer Vision and Pattern Recognition. Springer, Cham. https://doi.org/10.1007/978-3-319-08651-4_11.

[8] Herv´e J´egou, Matthijs Douze, Cordelia Schmid, and Patrick P´erez. Aggregating local descriptors into a compact image representation. In Computer Vision and Pattern Recognition Conference on, pages 3304–3311. IEEE, 2010.

[9] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. Neurocomputing, 70(1):489–501, 2006.

[10] Wang, R.Y.; Popovi´c, J. Real-time hand tracking with a color glove. ACM Trans. Graph. 2009, 28, 1–8.

[11]. Schroder, M.; Elbrechter, C.; Maycock, J.; Haschke, R.; Botsch, M.; Ritter, H. Real-time hand tracking with a color glove for the actuation of anthropomorphic robot hands. In Proceedings of the 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids), Osaka, Japan, 29 November–1 December 2012; pp. 262–269.

[12] Shotton, J.; Sharp, T.; Fitzgibbon, A.; Blake, A.; Cook, M.; Kipman, A.; Finocchio, M.; Moore, R. Real-Time human pose recognition in parts from single depth images. Commun. ACM 2013, 56, 116–124.

[13] Maryam Asadi-Aghbolaghi, Albert Clap´es, Marco Bellantonio, Hugo Jair Escalante, V´ıctor Ponce-L´opez, Xavier Bar´o, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera. Deep learning for action and gesture recognition in image sequences: A survey. In Gesture Recognition, pages 539–578. Springer, 2017.
[14] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4207–4215, 2016.
[15] Oyedotun, O.K.; Khashman, A. Deep learning in vision-based static hand gesture recognition. Neural Comput. Appl. 2017, 28, 3941–3951.
[16] Molchanov, P.; Gupta, S.; Kim, K.; Kautz, J. Hand gesture recognition with 3D convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern RecognitionWorkshops, Boston, MA, USA, 7–12 June 2015; pp. 1–7.
[17] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition. In NIPS, 2014.
[18]] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In ICCV, 2015.

Base Paper