Jyothi Hills, Panjal Road,
Vettikattiri PO, Cheruthuruthy , Thrissur,
Kerala 679531

Jyothi Engineering College
NAAC Accredited College with NBA Accredited Programmes*
Approved by AICTE & affiliated to APJ Abdul Kalam Technological University
A CENTRE OF EXCELLENCE IN SCIENCE & TECHNOLOGY BY THE CATHOLIC ARCHDIOCESE OF TRICHUR
JYOTHI HILLS, VETTIKATTIRI P.O, CHERUTHURUTHY, THRISSUR. PIN-679531 PH : +91- 4884-259000, 274423 FAX : 04884-274777
NBA accredited B.Tech Programmes in Computer Science & Engineering, Electronics & Communication Engineering, Electrical & Electronics Engineering and Mechanical Engineering valid for the academic years 2016-2022. NBA accredited B.Tech Programme in Civil Engineering valid for the academic years 2019-2022.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SEMINAR REPORT

Hand Gesture Recognition Using 3D Dynamic Skeletal Data

Submitted by

SANDRA DAVID
JEC17CS084

Supervised by

Ms. Ninu Francis
Asst. Prof., Dept. of CSE

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY (B.Tech)

in

COMPUTER SCIENCE & ENGINEERING
of

A P J ABDUL KALAM TECHNOLOGICAL UNIVERSITY

CREATING TECHNOLOGY LEADERS OF TOMORROW

DECEMBER 2020

Jyothi Hills, Panjal Road,
Vettikattiri PO, Cheruthuruthy , Thrissur,
Kerala 679531

Jyothi Engineering College
NAAC Accredited College with NBA Accredited Programmes*
Approved by AICTE & affiliated to APJ Abdul Kalam Technological University
A CENTRE OF EXCELLENCE IN SCIENCE & TECHNOLOGY BY THE CATHOLIC ARCHDIOCESE OF TRICHUR
JYOTHI HILLS, VETTIKATTIRI P.O, CHERUTHURUTHY, THRISSUR. PIN-679531 PH : +91- 4884-259000, 274423 FAX : 04884-274777
NBA accredited B.Tech Programmes in Computer Science & Engineering, Electronics & Communication Engineering, Electrical & Electronics Engineering and Mechanical Engineering valid for the academic years 2016-2022. NBA accredited B.Tech Programme in Civil Engineering valid for the academic years 2019-2022.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SEMINAR REPORT

Hand Gesture Recognition Using 3D Dynamic Skeletal Data

Submitted by

SANDRA DAVID
JEC17CS084

Supervised by

Ms. Ninu Francis
Assist. Prof., Dept. of CSE

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY (B.Tech)

in

COMPUTER SCIENCE & ENGINEERING
of

A P J ABDUL KALAM TECHNOLOGICAL UNIVERSITY

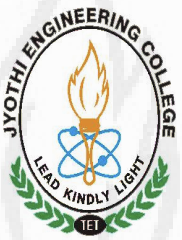
CREATING TECHNOLOGY LEADERS OF TOMORROW

DECEMBER 2020

Department of Computer Science and Engineering

JYOTHI ENGINEERING COLLEGE, CHERUTHURUTHY

THRISSUR 679 531



DECEMBER 2020

BONAFIDE CERTIFICATE

This is to certify that the seminar report entitled **Hand Gesture Recognition Using 3D Dynamic Skeletal Data** submitted by **Sandra David (JEC17CS084)** in partial fulfillment of the requirements for the award of **Bachelor of Technology** degree in **Computer Science and Engineering** of **A P J Abdul Kalam Technological University** is the bonafide work carried out by her under our supervision and guidance.

Ms. Ninu Francis	**Dr. Swapna B Sasi**	**Dr. Vinith R**
Seminar Guide	Seminar Coordinator	Head of The Dept.
Assistant Professor	Associate Professor	Professor
Dept. of CSE	Dept. of CSE	Dept. of CSE

DEPARTMENT OF

COMPUTER SCIENCE & ENGINEERING

COLLEGE VISION

Creating eminent and ethical leaders through quality professional education with emphasis on holistic excellence.

COLLEGE MISSION

- To emerge as an institution par excellence of global standards by imparting quality engineering and other professional programmes with state-of- the-art facilities.
- To equip the students with appropriate skills for a meaningful career in the global scenario.
- To inculcate ethical values among students and ignite their passion for holistic excellence through social initiatives.
- To participate in the development of society through technology incubation, entrepreneurship and industry interaction.

DEPARTMENT OF

COMPUTER SCIENCE & ENGINEERING

DEPARTMENT VISION

Creating eminent and ethical leaders in the domain of computational sciences through quality professional education with a focus on holistic learning and excellence.

DEPARTMENT MISSION

- To create technically competent and ethically conscious graduates in the field of Computer Science & Engineering by encouraging holistic learning and excellence.
- To prepare students for careers in Industry, Academia and the Government.
- To instill Entrepreneurial Orientation and research motivation among the students of the department.
- To emerge as a leader in education in the region by encouraging teaching, learning, industry and societal connect

PROGRAMME OUTCOMES (POs)

1. **Engineering Knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem Analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/Development of Solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct Investigations of Complex Problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
5. **Modern Tool Usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The Engineer and Society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and Sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and Team Work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project Management and Finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-Long Learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAMME EDUCATIONAL OBJECTIVES (PEOs)

1. The graduates shall have sound knowledge of Mathematics, Science, Engineering and Management to be able to offer practical software and hardware solutions for the problems of industry and society at large.
2. The grad uates shall be able to establish themselves as practising professionals, researchers or Entrepreneurs in computer science or allied areas and shall also be able to pursue higher education in reputed institutes.
3. The graduates shall be able to communicate effectively and work in multidisciplinary teams with team spirit demonstrating value driven and ethical leadership.

Programme Specific Outcomes (PSOs)

1. An ability to apply knowledge of data structures and algorithms appropriate to computational problems.
2. An ability to apply knowledge of operating systems, programming languages, data management, or networking principles to computational assignments.
3. An ability to apply design, development, maintenance or evaluation of software engineering principles in the construction of computer and software systems of varying complexity and quality.
4. An ability to understand concepts involved in modeling and design of computer science applications in a way that demonstrates comprehension of the fundamentals and tradeoffs involved in design choices.

Course Outcomes (COs)

C418.1 **Presentation Skills in terms of Content** : Students will be able to show competence in identifying relevant information, defining and explaining topics under discussion. They will demonstrate depth of understanding, use primary and secondary sources; they will demonstrate the working, complexity, insight, cogency, independent thought, relevance, and persuasiveness. They will be able to evaluate information and use and apply relevant theories.

C418.2 **Presentation Skills in terms of Organization** : Students will be able to show competence in working with a methodology, structuring their oral work, and synthesizing information. They will make a detailed study on the previous works related to their topic and will present the observations.

C418.3 **Presentation Skills in terms of Delivery** : Students will use appropriate registers and vocabulary, and will demonstrate command of voice modulation, voice projection, and pacing. They will be able to make use of visual, audio and audio-visual material to support their presentation, and will be able to speak cogently with or without notes.

C418.4 **Discussion Skills** : Students will be able to judge when to speak and how much to say, speak clearly and audibly in a manner appropriate to the subject, ask appropriate questions, use evidence to support claims, respond to a range of questions, take part in meaningful discussion to reach a shared understanding, speak with or without notes, show depth of understanding.

C418.5 **Listening Skills** : Students will demonstrate that they have paid close attention to what others say and can respond constructively. Through listening attentively, they will be able to build on discussion fruitfully, supporting and connecting with other discussants.

C418.6 **Argumentative Skills and Critical Thinking** : Students will develop persuasive speech, present information in a compelling, well-structured, and logical sequence, respond respectfully to opposing ideas, show depth of knowledge of complex subjects, and develop their ability to synthesize, evaluate and reflect on information.

Course Outcome							
Programme Outcomes		C418.1	C418.2	C418.3	C418.4	C418.5	C418.6
	1	3	3	3	3	3	3
	2	3	3	3	3	3	3
	3	3	3	3	3	3	3
	4	3	3	3	3	3	3
	5	3	3	3	3	3	3
	6	3	3	3	3	3	3
	7	3	3	3	3	3	3
	8	3	3	3	3	3	3
	9	3	3	3	3	3	3
	10	3	3	3	3	3	3
	11	3	3	3	3	3	3
	12	3	3	3	3	3	3

PO - CO Mapping

PEO - CO Mapping

Course Outcome							
Programme Educational Objective		C418.1	C418.2	C418.3	C418.4	C418.5	C418.6
	1	3	3	1	1	-	2
	2	3	3	3	3	1	3
	3	1	2	3	3	1	3

PSO - CO Mapping

Course Outcome							
Programme Specific Outcomes		C418.1	C418.2	C418.3	C418.4	C418.5	C418.6
	1	3	3	3	3	3	3
	2	3	3	3	3	3	3
	3	3	3	3	3	3	3
	4	3	3	3	3	3	3

Seminar Outcome

1. Explored about the concept of Machine Learning.
2. Analyzed about the hand gesture recognition using Machine Learning.
3. Analyzed about the possibility of hand gesture recognition using 3D dynamic skeletal hand data.
4. Studied about Fisher Vector Representation and Temporal Pyramid.
5. Analyzed about SVM classification method.

Seminar Outcome - CO Mapping

Course Outcome							
Seminar Outcome		C418.1	C418.2	C418.3	C418.4	C418.5	C418.6
	1	3	3	3	1	3	3
	2	3	1	3	3	3	3
	3	3	3	3	1	3	3
	4	3	1	3	1	3	1
	5	3	3	3	3	3	1

ACKNOWLEDGEMENT

I take this opportunity to express my heartfelt gratitude to all respected personalities who had guided, inspired and helped me in the successful completion of this seminar. First and foremost, I express my thanks to **The Lord Almighty** for guiding me in this endeavour and making it a success.

I take immense pleasure in thanking the **Management** of Jyothi Engineering College and **Dr. Sunny Joseph Kalayathankal**, Principal, Jyothi Engineering College for having permitted me to carry out this seminar. My sincere thanks to **Dr. Vinith R**, Head of the Department of Computer Science and Engineering for permitting me to make use of the facilities available in the department to carry out the seminar successfully.

I express my sincere gratitude to **Dr. Swapna B Sasi & Mr. Shaiju Paul**, Seminar Coordinators for their invaluable supervision and timely suggestions. I am very happy to express my deepest gratitude to my mentor **Ms. Ninu Francis**, Assistant Professor, Department of Computer Science and Engineering, Jyothi Engineering College for her able guidance and continuous encouragement.

Last but not least, I extend my gratefulness to all teaching and non-teaching staff who directly or indirectly involved in the successful completion of this seminar work and to all friends who have patiently extended all sorts of help for accomplishing this undertaking.

ABSTRACT

Hand gestures come to us naturally. These are the easiest and quickest way of non-verbal communication. The hand gesture communication is effective for both human to human interaction as well as human-computer interaction (HCI). Hence, the field of hand gesture recognition has great relevance. Hand gesture recognition has gradually become an active field of research. The field of study opens up several applications in different aspects of the life. Thus, hand gesture recognition systems urged popularity, as an inspiring field of research. A number of papers have been published and a variety of approaches have been proposed, regarding the subject. Advances in the technology have promoted and improved the approach of 3D hand gesture detection and recognition. The report aims to explore the possibility of hand gesture recognition using 3D dynamic skeletal data of hand, with reference to the existing methodologies and papers. The approach relies on the structure of hand topology to extract the effective descriptors from the gesture sequence. The statistical representation method, Fisher Vector representation and the temporal representation method, temporal pyramid are used for the encoding of the descriptors. Finally, an SVM classifier is used for the classification of the gesture. The report evaluates the already proposed methodology of hand gesture recognition using 3D dynamic skeletal data.

CONTENTS

List of Figures

List of Abbreviations

HCI	: *Human Computer Interaction*
FV	: *Fisher Vector*
TP	: *Temporal Pyramid*
SVM	: *Support Vector Machine*
LMC	: *Leap Motion Controller*
CNN	: *Convolutional Neural Network*

CHAPTER 1
INTRODUCTION

1.1 Overview

A hand gesture is an intuitive medium of non-verbal communication. It is an effective natural means of communication. There exists a number of standard and accepted hand gestures worldwide. Many hand gestures have different meanings in different countries, in fact their meanings may even be contradictory from place to place.

Hand gesture recognition has several applications. The fields of applications are wide and varied. The applications range from human-human communication to human-computer communication. It has applications in virtual environment control, sign language recognition, robot control, etc. Hence, it is an area of research with a wide scope. Hand gesture recognition has been an active research field for the past 20 years. Over the past few years, the advances in technology, like the 3D depth sensors, has boosted the hand gesture detection and recognition field.

The report explores the existing method of hand gesture recognition using 3D dynamic skeletal data. The approach relies on the structure of hand topology to extract the effective descriptors from the gesture sequence. However, the hand is an object with a complex topology.The hand forms merely a small portion of our body. Also there are many concerns like the endless possibilities of doing the same gesture, the heterogeneity among the hands in data set, etc. that have to be tackled wisely and efficiently.The recently popular devices such as, Intel Real Sense or Leap Motion Controller provides precise skeletal data of hand with 22 joints. Each additional methodology you add in the procedure of the approach, the accurate is the expected result. The statistical representation method, Fisher Vector representation and the temporal representation method, temporal pyramid are used for the encoding of the descriptors. Finally, an SVM classifier is used for the classification of the gesture.

1.2 Objective

The main objective of the paper is to analyse the existing method of the detection and recognition of hand gestures using 3D dynamic skeletal data of hand. The hand gesture is

considered as a time series of hand skeletons. It takes into consideration the direction of hand, its rotation and its shape during the sequence of gesture.

1.3 Organization Of The Report

The report is organised as follow:

- **Chapter 1:Introduction** The chapter gives an introduction to the hand gesture detection and recognition system using 3D dynamic skeletal data.
- **Chapter 2: Literature Survey** The chapter summarizes the researches on different systems for the gesture detection and recognition.
- **Chapter 3: Approach**The chapter discusses in depth about an approach for hand gesture recognition system using 3D dynamic skeletal data step wise.
- **Chapter 4: Implementation & Results** The chapter contains the implementation and results of the approach.
- **Chapter 5: Advantages & Applications** The chapter lists out the advantages and of using hand skeletal data, FV representation and TP representation for hand gesture recognition system. Also, lists the applications of the recognition system.
- **Chapter 6: Conclusion** The overall development and its inferred results are concluded.
- **References** Includes the references of the papers related to gesture detection systems and the methodologies used.

CHAPTER 2
LITERATURE SURVEY

2.1 Skeleton-Based Dynamic Hand Gesture Recognition

In the paper a new approach for 3D hand gesture recognition is proposed. The Intel Real Sense camera is used for the information of hand joints. The geometric shape of hand is used for the extraction of features. Each descriptor is then encoded by a Fisher Vector representation obtained using a Gaussian Mixture Model. The task is to evaluate a challenging dataset of 14 hand gestures. The main approach of the approach is the use of hand skeleton data. The experiments proved that the the skeleton based approach is consistent for the gesture recognition system.[1]

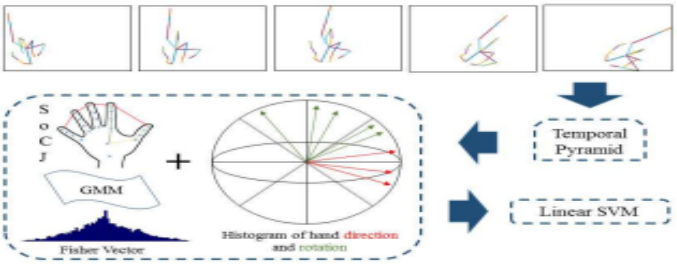


Figure 2.1: Pipeline of the proposed system

2.1.1 Feature Extraction

Feature extraction is the process of reducing dimension by which an initial set of raw data is reduced to more manageable groups for processing. Here it is made sure that not just the hand shape is considered but also the movement and rotation. The final classification is performed by a linear SVM classifier.

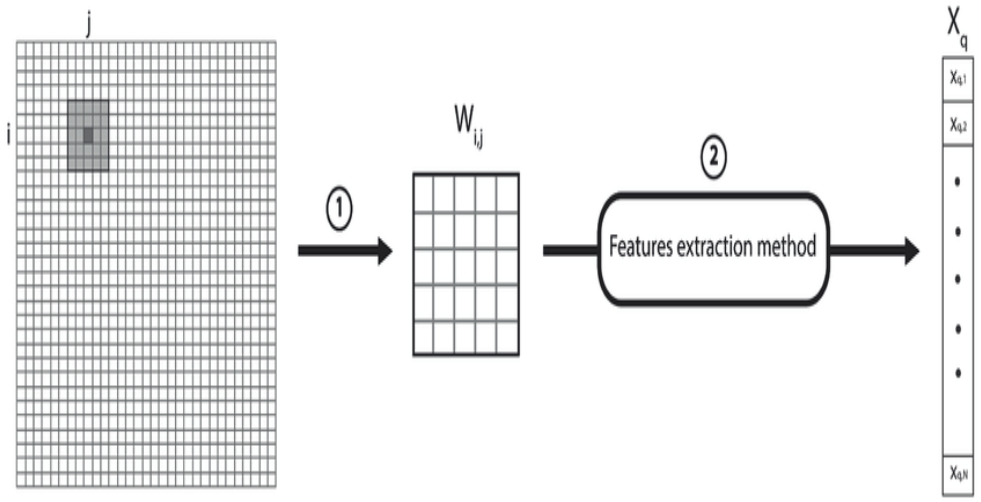


Figure 2.2: Feature extraction

Shape of Connected Joints

To represent hand using hand skeleton, a descriptor based on sets of joints is defined. Hand skeleton returned by the sensors contain 3D coordinates of hand joints. This creates a variance between the hand of performers in camera. To compensate this variance, the average size of each bone is estimated. Now keeping the angles unchanged, change size of hand by the mean found previously. Now assume a fake hand with its palm open to the camera and with joints [0 0 0]. Let this be the base hand Hf for reference.

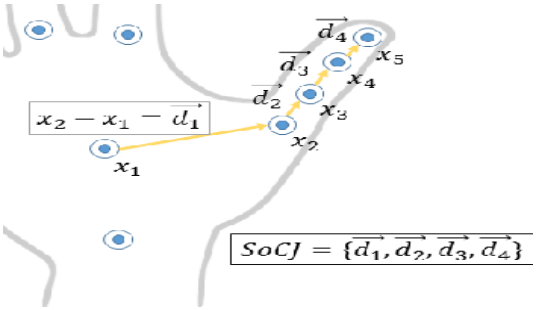


Figure 2.3: An example of the SoCJ descriptor

The translation and rotational estimations of the first image of the sequence is estimated. This estimations are applied to all other images of the sequence. To estimate the SoCJ we find the difference between the selected tuple of joints.

Fisher Vector Representation

FV method was introduced for the large scale image classification. It encodes the information of descriptors. This method is superior to the BOW method. But before performing the FV, it is to be noted that a GMM is to be constructed.

Other relevant features

Histogram of Hand Directions(HoHD): Some gestures can be defined only by the way how the hand moves in space. A set to define the direction of hand movement is defined.

Histogram of Wrist Rotations(HoWR): In some cases the rotation of wrist plays an important role. It is important to define a set to represent the rotation of wrist.

Temporal Modelling and Classification

To add the approach of temporal representation, the simple representation method called Temporal Pyramid is used. The principal is to divide the sequence into n levels with each level i having i sub sequences.

2.1.2 Conclusion

The work suggests the advantage of using 3D hand skeletal data to describe hand gestures. The input is several sets of relevant joints inferred from the hand skeleton. The approach assures an accuracy of 83 per-cent from experiments.

2.2 Robust 3D Hand Pose Estimation in Single Depth Images: from Single-View CNN to Multi-View CNNs

The hand pose estimation has an important role in human computer interaction. The work suggests the accuracy of the existing methods are not satisfactory. The data-driven methods for hand pose estimation train discriminative models, such as isometric self-organizing map [6] and convolution neural networks (CNNs) [15], to map the image features to the hand pose parameters.

2.2.1 Methodology

The task is the extraction of the 3D hand joint locations from the depth image. . The input of the work is a cropped depth image only containing a human hand with some gesture and the outputs are K 3D hand joint locations which represent the hand pose. The 21 objective hand joint locations are the wrist center, the five metacarpophalangeal joints, the five proximal interphalangeal joints, the five distal interphalangeal joints and the five finger tips.[5] A multiview projection and learning method is followed. The objective is to generate projected images. The details of 3D projections are described. Then the architecture of CNNs is introduced.

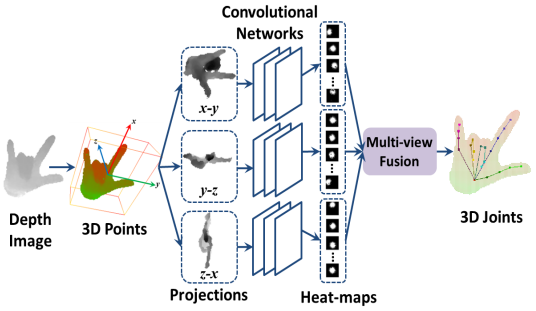


Figure 2.4: Overview of the proposed framework

3D Points Projection

The input depth image is converted to a set of 3D points. These points are projected to orthogonal planes.The 3D points are projected to three views. For each view, Convolutional network having same architecture are constructed. The distances from these points to the plane have to be normalized from 0 to 1.

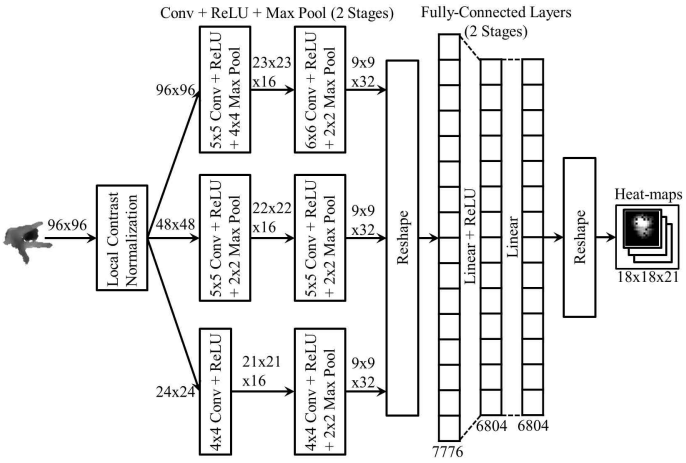


Figure 2.5: Convolutional Network architecture

Multi-view fusion

The purpose of the multi-view fusion is to estimate the 3D hand joint locations from three views' heat-maps. The CNNs generate a set of heat-maps for each joint, each view. Since the intensity on a heat-map indicates the confidence of a joint locating in the 2D position of the x-y, y-z or z-x view, we can get the corresponding probabilities.[15]To simplify the problem, the product of probabilities, is approximated as a 3D Gaussian distribution.

Self comparison

A self comparison strategy is to be employed in the proposed system. The two baselines implemented are, the single view regression approach and the multi-view regression approach using a coarse fusion method. In the single view regression approach, only the projected images on OBB coordinate system's x-y plane are fed into the CNNs. . The multi-view regression approach using a coarse fusion method can be considered as a degenerated variant of our fine fusion method. This method estimates the 3D hand joint locations by simply averaging the estimated x, y and z coordinates from three views' heat-maps.[15]

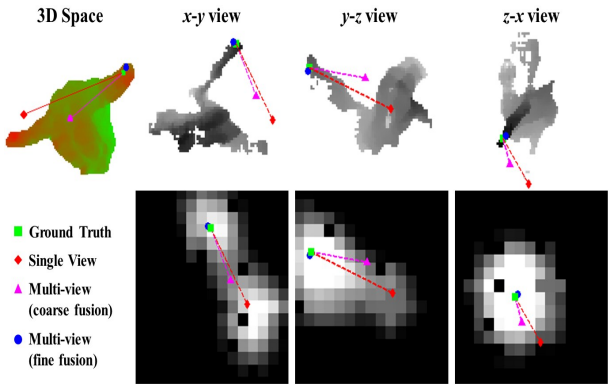


Figure 2.6: An experimental example for self comparison

Cross-dataset Experiment

The generalization ability of the multi-view regression using CNN can be verified by performing a cross-dataset experiment. The CNNs are directly used for hand pose estimation. The process of multi-view projection and multi-view fusion are performed on CPU without parallelism, and the process of CNNs forward propagation is performed on GPU with parallelism for three views.

2.2.2 Conclusion

In the work, a 3D hand pose regression method using multi-view CNNs is proposed. The work gives better information about the 3D joints. The paper claims superior performance of hand pose estimation.

2.3 Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks

An automatic system for detection and classification of dynamic hand gestures in real world is challenging. In the proposed work, it is hoped to address these challenges with a recurrent three-dimensional convolutional neural network that performs simultaneous detection and classification of dynamic hand gestures from multi-modal data. A temporal classification is used to train the network to predict the class labels. The accuracy of the recognition system claimed is 83.8 per-cent.[11]

Hand gesture recognition has several important applications. It is important in several instances such as to improve comfort and safety in vehicles. But the real world systems posses various challenges. These systems receive continuous streams of unprocessed visual data. The dynamic hand gestures generally contain three temporally overlapping phases: preparation, nucleus, and retraction [4, 9], of which the nucleus is most discriminative.

2.3.1 Method

Network Architecture

A recurrent 3D convolutional neural network (R3DCNN) for dynamic hand gesture recognition is proposed. The operations performed by the network are to be formalized. The multiple modalities by averaging the class conditional probabilities are to be estimated by the modality-specific networks.

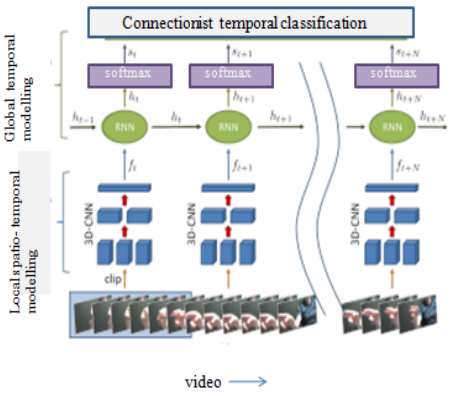


Figure 2.7: Classification of dynamic gestures with R3DCNN

Training

Pre-training the 3D-CNN: The 3D-CNN is initialized with the C3D network. The network posses 8 convolutional layers of 3×3×3 filters and 2 fully-connected layers trained on 16-frame clips. A softmax prediction layer is appended to the last fully-connected layer.

Training the full model: The entire R3DCNN is trained with back-propagationthrough-time (BPTT). This is to unroll the recurrent layers, transforming them to a multi-layer feed-forward network. Two training cost functions: negative loglikelihood for the entire video and connectionist temporal classification (CTC) for online sequences are considered.

Connectionist temporal classification: CTC is a cost function that is designed for sequence prediction. In the work, CTC is employed to identify and correctly label the nucleus of gesture. The system not only classifies and labels a gesture but also identifies a 'no gesture'. Hence, it also requires a no gesture class. Instead of averaging the predictions, the network computes the probability of observing a particular gesture (or no gesture). A path is defined, as a possible mapping of the input sequence X into a sequence of class labels y. The computation of p(y|X) is simplified by dynamic programming. For pre-segmented video classification, simply remove the no-gesture output and re-normalize probabilities.

Regularization

The purpose of the regularization method is to reduce overfitting. For this, train with weight decay (= 0.5per-cent) on all weights in the network. Then apply drop-out to the fully-connected layers of the 3D-CNN at a rate of p= 75per-cent, rescaling the remaining activations by a factor of 1/(1 p). al layers improves generalization in pre-trained networks. Finally,

randomly set 10per-cent of the feature maps of each convolutional layer to 0 and rescale the activations of the others neurons accordingly.

Figure 2.8: Environment for data collection

2.3.2 Results

Sensors	Accuracy	Combinations									
Depth	80.3%			✓	✓		✓	✓	✓	✓	✓
Optical flow	77.8%		✓	✓		✓	✓	✓	✓	✓	✓
Color	74.1%		✓	✓	✓	✓				✓	✓
IR image	63.5%	✓			✓	✓		✓	✓	✓	✓
IR disparity	57.8%	✓							✓		✓
Fusion Accuracy		66.2%	79.3%	81.5%	82.0%	82.0%	82.4%	82.6%	83.2%	83.4%	**83.8%**

Figure 2.9: Comparison of modalities and their combinations

2.3.3 Conclusion

The work proposed a novel recurrent 3D convolutional neural network classifier for dynamic gesture recognition. It supports online gesture classification with zero or negative lag, effective modality fusion, and training with weakly segmented videos. The work assured better accuracy with respect to the other proposed systems.

2.4 Deep Residual Learning for Image Recognition

A residual learning framework to ease the training of networks that are substantially deeper than those used previously is proposed in the work. Deep convolutional networks have showcased an important role in image classification. The work proposes to explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. The idea is to provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth.When deeper networks are able to start converging, a degradation problem has been exposed: with the network depth increasing, accuracy gets saturated (which might be unsurprising) and then degrades rapidly.[7] The degradation problem is solved by deep residual learning framework. The work proposes to explicitly let the stacked layers fit a residual mapping, rather than hoping each few stacked layer.

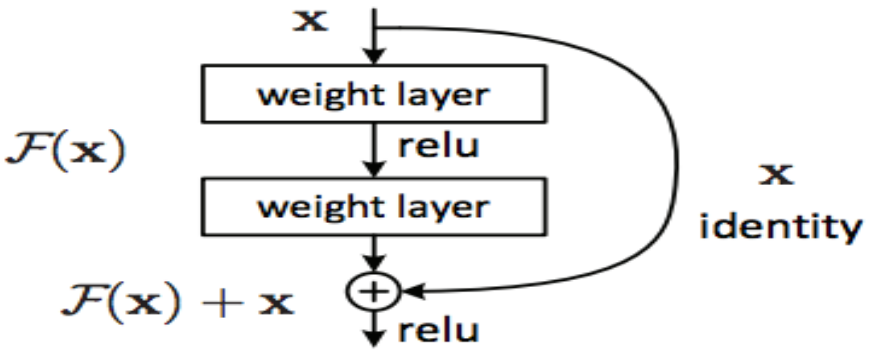


Figure 2.10: Residual learning

2.4.1 Deep Residual Learning

Residual Learning

The work proposes to explicitly let the stacked layers approximate a residual function F(x) := H(x) x. The degradation problem suggests that the solvers might have difficulties in approximating identity mappings by multiple nonlinear layers. With the residual learning reformulation, if identity mappings are optimal, the solvers may simply drive the weights of the multiple nonlinear layers toward zero to approach identity mappings.[7] The experiments showed that the learned residual functions in general have small responses, suggesting that identity mappings provide reasonable preconditioning.

Identity Mapping by Shortcuts

The residual learning is adopted to every few stacked layers. The form of the residual function F is flexible. The operation F + x is performed by a shortcut connection and element-wise addition. This is important in comparison between the plain and residual networks. Experiments in the paper involve a function F that has two or three layers (Fig. 2.10), while more layers are possible.

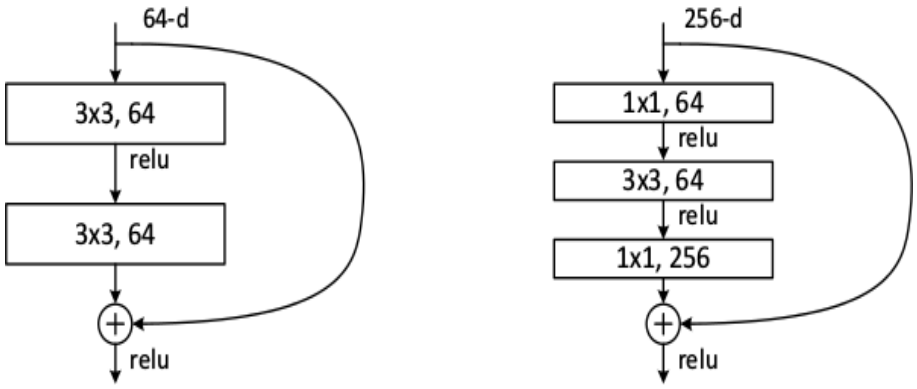


Figure 2.11: A deeper residual function F for ImageNet. Left: a building block (on 56×56 feature maps) for ResNet34. Right: a "bottleneck" building block for ResNet-50/101/152

2.4.2 Network Architectures

Through the experiments various plain/residual nets were tested and consistent phenomena were observed. To provide instances for discussion, the describe two models for ImageNet as follows.

Plain Network

The model has fewer filters and lower complexity than VGG nets. The convolutional layers mostly have 3×3 filters and follow two simple design rules: (i) for the same output feature map size, the layers have the same number of filters; and (ii) if the feature map size is halved, the number of filters is doubled so as to preserve the time complexity per layer.[7]

Residual Network

Some shortcut connections which turn the network into its counterpart residual version, are inserted to plain network. The shortcut performs identity mapping. But the identity shortcuts

are performed only when the input and output are of same dimensions.

2.4.3 Implementation

The implementation for ImageNet follows the practice in [10, 14]. The image is resized with its shorter side. The method adopts a batch normalization (BN) right after each convolution and before activation, following. Then innitialize the weights as in and train all plain/residual nets from scratch. . The learning rate starts from 0.1. The standard 10-crop testing is used for comparison. A fully convolutional form is adopted and average the scores at multiple scales.

2.4.4 CIFAR-10 and Analysis

Studies on CIFAR-10 are conducted.aining images and 10k testing images in 10 classes. Here, experiments trained on the training set and evaluated on the test set are presented. The focus is on the deep networks. The plain/residual architectures follow the form in Fig. 3 (middle/right). The plain/residual architecture network inputs are 32×32 images, with the per-pixel mean subtracted. The use of a weight decay of 0.0001 and momentum of 0.9 is preferred, and adopt a weight initialization and BN but with no dropout. These models are trained with a minibatch size of 128 on two GPUs. When there are more layers, an individual layer of ResNets tends to modify the signal less.

2.5 Action and Event Recognition with Fisher Vectors on a Compact Feature Set

Action recognition in uncontrolled video is an important and challenging computer vision problem. Instead of working towards more complex models, the focus is on the low-level features and their encoding. The evaluation is using Fisher vectors as an alternative to bag-of-word histograms to aggregate a small set of state-of-the-art low-level descriptors, in combination with linear classifiers. [12]

The recognition from video posses the challenges due to sheer amount of data that need to be processed. In the paper, the potential of the Fisher vector (FV) encoding as a robust feature pooling technique that has proven to be among the most effective for object recognition is explored. First, consider the classification of basic action categories using five of the most challenging recent datasets. Second, consider the localization of actions in feature length movies, using the four actions drinking, smoking, sit down, and open door.[12]

2.5.1 Video Representation

Feature Extraction

The first task is to encode the low level visual content using static appearance features as well as motion features. Now compute SIFT descriptors every tenth video frame. Then capture motion information using the recently introduced dense trajectory Motion Boundary Histogram (MBH). The MBH feature is similar to SIFT, but computes gradient orientation histograms over both the vertical and horizontal spatial derivatives of the optical flow.

Feature Encoding

Once the two local low-level features sets are extracted, use them to construct a signature to characterize the video. For this step we use the Fisher vector (FV) representation, which was found to be the most effective one in a recent evaluation study of feature pooling techniques for object recognition, which included FVs, bag-of-words. The FV records, for each quantization cell, not only the number of assigned descriptors, but also their mean and variance along each dimension. Instead of using k-means clustering, Gaussian mixture clustering is used in the FV representation.

Weak spatio-temporal location information

To go beyond a completely orderless representation of the video content in a single FV, consider including a weak notion of spatio-temporal location information of the local features. For this purpose, use the spatial pyramid (SPM) representation, and compute separate FVs over cells in spatio-temporal grids. It is important to also consider the spatial Fisher vector (SFV).

Non-maximum-suppression for localization

For the action localization task employ a temporal sliding window approach. Then score a large pool of candidate detections that are obtained by sliding windows of various lengths across the video. . Longer windows might better cover the action, but are likely to include less characteristic features (even if they lead to positive classification by themselves), and might include background features due to imperfect temporal alignment. To address this issue we consider re-scoring the segments by multiplying their score with their duration, before applying NMS (referred to as RS-NMS). We also consider a variant where the goal is to select a subset of candidate windows that (i) covers the video, (ii) does not have overlapping windows, and

(iii) maximizes the sum of scores of the selected windows. [12]

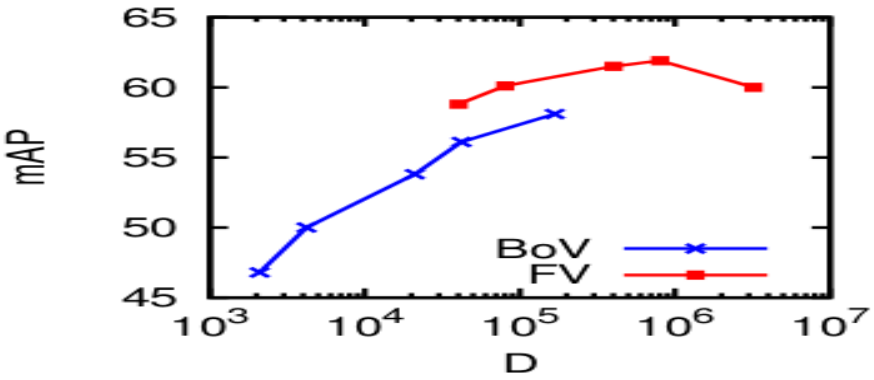


Figure 2.12: Event recognition with FV

2.5.2 Experiment

In the first step, we only use the MBH descriptor, and compare FV and BOV. The experiments show that FVs using 50 visual words are comparable to BoV histograms for 4000 visual word. On all datasets our performance is comparable or better than the current state of the art using only MBH features. The SIFT features perform significantly worse, and carry relatively little useful complementary information. In our second set of experiments we consider the localization of four actions in feature length movies. Given the size of the test dataset, we encode both the MBH and SIFT features with FVs with K = 128 Gaussians and do not include location information with SPM or SFV. [12] The last set of experiments is to consider the TrecVid MED 2011 event recognition dataset. For both features we use K = 256 visual words, and exclude SPM and SFV for efficiency. Comparing the results with the previously existing reports is important for performance.

2.5.3 Conclusion

The proposed system claims to be an efficient action recognition system that combines three state-of-the-art low-level descriptors (MBH,SIFT, MFCC) with the recent Fisher vector representation. In the experiment, the thing considered were the action recognition, action localization in movies, and complex event recognition. The evaluation is among the most extensive and diverse ones to date, including five of the most challenging action recognition benchmarks, action localization in feature length movies, and large-scale event recognition.

2.6 A Framework for Recognizing the Simultaneous Aspects of American Sign Language

The major challenge that faces American Sign Language (ASL) recognition now is developing methods that will scale well with increasing vocabulary size. The number of possible combinations of phonemes is approximately 1.5 × 109, which cannot be tackled by conventional hidden Markov model-based methods. Sign language recognition enters the picture in three ways. First, such a paradigm shift would leave those deaf people who depend on sign language as their primary mode of communication behind. There is a sense of urgency, because of the improvements in speech recognition. Unless we get sign language recognition to the same level of performance as speech recognition, accessibility of computers will become a major issue for the deaf. Second, gesture recognition in itself is a difficult problem, because gestures are unconstrained, but gestures take place in the same visual medium as sign languages, and the latter possess a high degree of structure. This structure makes it easier to solve problems in sign language recognition first, before applying them to gesture recognition. Third, a working sign language recognition system would make deaf–hearing interaction easier. Particularly public functions, such as the courtroom, conventions, and meetings, would become much more accessible to the deaf. [17]

The main challenge in sign language recognition is to find a modeling paradigm that is powerful enough to capture the language. In the paper, a novel framework for modeling and recognizing American Sign Language (ASL) is presented. A 3D data is used as an input for the recognition framework.

2.6.1 Modelling ASL

In USA, most deaf people use ASL to communicate. The ideas behind ASL phonology can be broken into small parts. There are one-handed signs and two-handed signs. The one-handed signs and the major component of the two-handed signs are performed by the dominant hand or strong hand. For right-handed people, the strong hand is the right hand and for left-handed people it is the left hand.

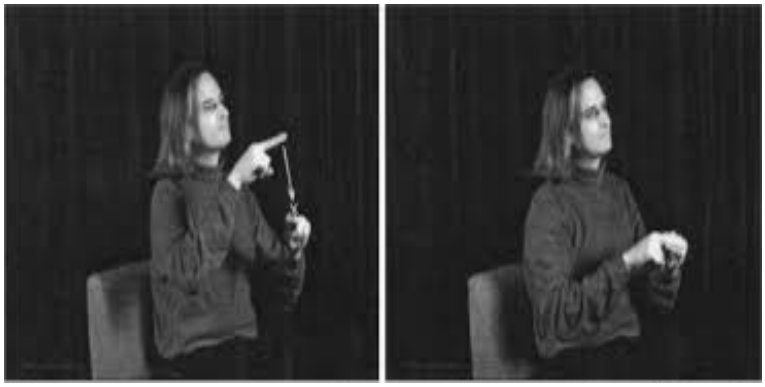


Figure 2.13: A random example for a sign language

The Movement–Hold Model

Liddell and Johnson's Movement–Hold model is one of the oldest segmental models. It consists of two major classes of segments, which are movements and holds. Signs are sequences of movements. In this paper we use only the aspects of the Movement–Hold model that describe hand movements and locations, because these are the easiest to capture with our 3D tracking system. The locations can be modified with distances from body. Holds, are those segments, during which the hands remain translationally stationary. The Movement–Hold model does not address nonmanual features, such as facial expressions. Because facial expressions constitute a large part of the grammar of signed languages, future work needs to address this shortcoming.

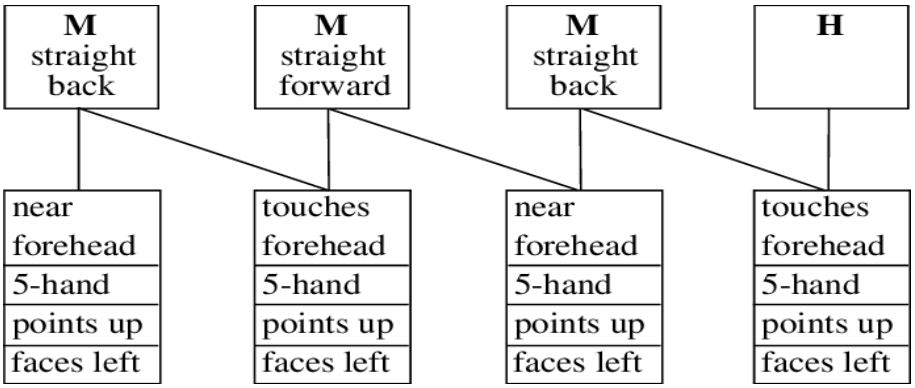


Figure 2.14: MMMH pattern. The sign for "father" consists of three movements: tap on forehead, away from forehead, tap on forehead (left), followed by a hold contacting the forehead (right)

2.7 Conclusion

Semantic representation of ASL will also be important, particularly for deaf–hearing interaction. Because the structure of ASL is so different from spoken languages, it is necessary to do more research into parsing the recognized ASL constructs and converting them into a semantic representation.

2.8 Temporal Pyramid Network for Action Recognition

Visual tempo characterizes the dynamics and the temporal scale of an action. Modeling such visual tempos of different actions facilitates their recognition.[20]The visual tempo describes how an action goes. The two essential components of TPN, the source of features and the fusion of features, form a feature hierarchy for the backbone so that it can capture action instancesat various tempos. In the work, a general formulation of the proposed TPN, where several components are introduced to better capture the information at multiple visual tempos is proposed. The TPNs are evaluated on three benchmarks.

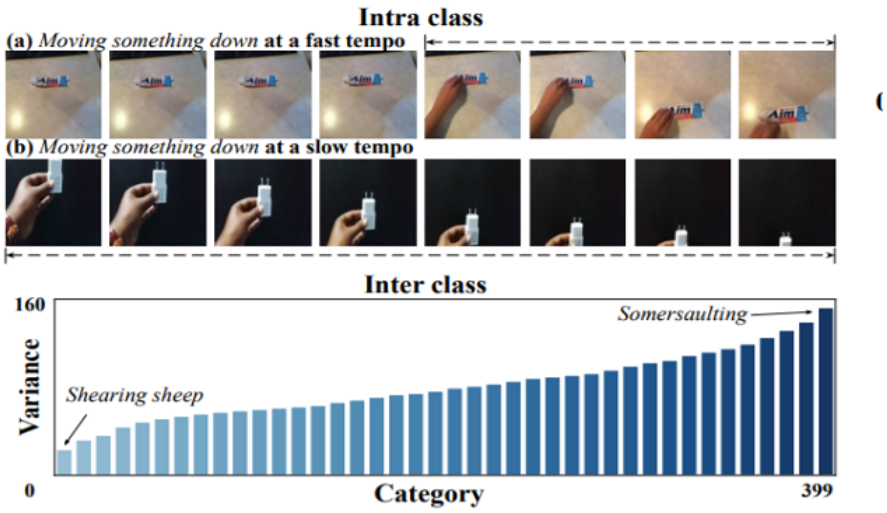


Figure 2.15: Visual tempo variation of intra- and inter-class

2.8.1 Temporal Pyramid Network

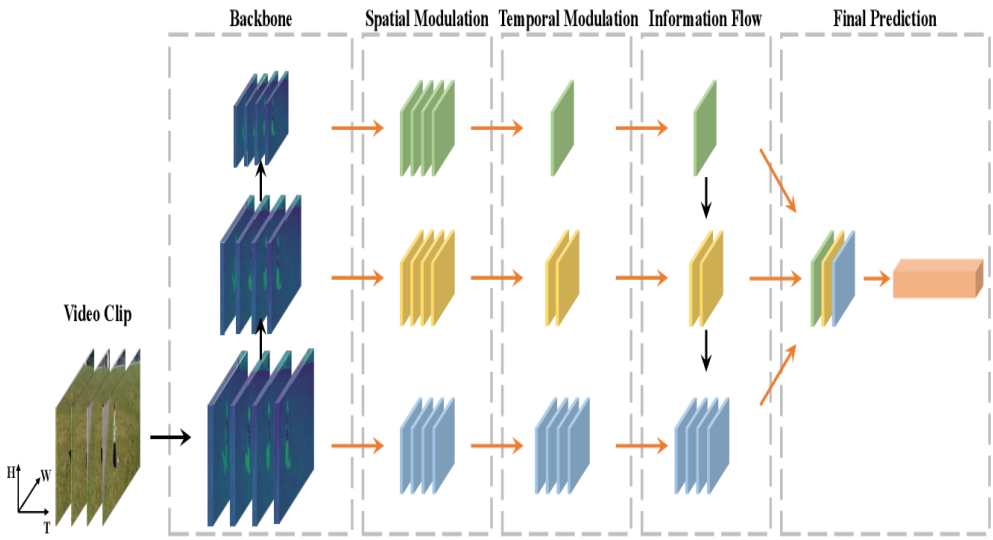


Figure 2.16: Framework of TPN

Inspired by the observation that features at multipledepths in a single network already cover various visual tempos, we propose a feature-level temporal pyramid network (TPN) for modeling the visual tempo. TPN could operate on only a single network no matter how many levels are included in it. Moreover, TPN could be applied to different architectures in a plug-and-play manner. To fully implement TPN, two essential components of TPN must be designed properly, namely 1) the feature source and 2) the feature aggregation. The work propose the spatial semantic modulation and temporal tempo modulation to control the relative differences of the feature source.[20]

2.8.2 Implementation

The proposed TPN is evaluated on various action recognition datasets. The features of TPN will be separately rescaled by max-pooling operations, and their concatenation will be fed into a fully-connected layer to make the final predictions. TPN can be also jointly trained with the backbone network in an end-to-end manner. Several empirical analysis are presented.

Backbone

We evaluate TPN on both 2D and 3D backbone networks. Specifically, the slow-only branch of SlowFastis applied as our backbone network (denoted as I3D) due to its promising performance on various datasets.

Stage	Layer	Output size
raw	-	$8 \times 224 \times 224$
conv_1	$1\times7\times7$, 64, stride 1, 2, 2	$8\times112\times112$
pool_1	$1\times3\times3$ max, stride 1, 2, 2	$8\times56\times56$
res_2	$\begin{bmatrix} 1\times1\times1, 64 \\ 1\times3\times3, 64 \\ 1\times1\times1, 256 \end{bmatrix} \times3$	$8\times56\times56$
res_3	$\begin{bmatrix} 1\times1\times1, 128 \\ 1\times3\times3, 128 \\ 1\times1\times1, 512 \end{bmatrix} \times4$	$8\times28\times28$
res_4	$\begin{bmatrix} 3\times1\times1, 256 \\ 1\times3\times3, 256 \\ 1\times1\times1, 1024 \end{bmatrix} \times6$	$8\times14\times14$
res_5	$\begin{bmatrix} 3\times1\times1, 512 \\ 1\times3\times3, 512 \\ 1\times1\times1, 2048 \end{bmatrix} \times3$	$8\times7\times7$
global average pool, fc		$1\times1\times1$

Figure 2.17: 3D Backbone

Results on Something-Something: For a fair comparison, we use the center crop of size 224×224 in all 8 segments.

Results on Epic-Kitchen: Here compare TSN+TPN to two baselines on Epic-Kitchen.

2.8.3 Conclusion

TPN with other state-of-the-art methods on Kinetics-400. In the paper, a generic module called Temporal Pyramid Network is proposed to capture the visual tempos of action instances. Our TPN, as a feature-level pyramid, can be applied to existing 2D/3D architectures in the plug-andplay manner, bringing consistent improvement

Model	Frames	Flow	Top-1	Top-5
R(2+1)D [28]	16	✓	73.9	90.9
I3D [1]	16	✓	71.6	90.0
Two-Stream I3D [1]	64	✓	75.7	92.0
S3D-G [34]	64	✓	77.2	93.0
STC-X101 [4]	32		68.7	88.5
Nonlocal-R50 [32]	32		76.5	92.6
Nonlocal-R101 [32]	32		77.7	93.3
SlowFast-R50 [5]	32		77.0	92.6
SlowFast-R101 [5]	32		77.9	93.2
CSN-101 [27]	32		76.7	92.3
CSN-152 [27]	32		77.8	92.8
TPN-R50	32×2		77.7	93.3
TPN-R101	32×2		**78.9**	**93.9**

Figure 2.18: Comparison with other state-of-the-art methods on the validation set of Kinetics-400

CHAPTER 3

HAND GESTURE RECOGNITION USING 3D DYNAMIC SKELETAL DATA

The hand occupies relatively a small portion of our body, but has a complex topology. There are endless possibilities for performing a gesture. Some gestures are identified by the hand shape whereas, some other by the hand motion. In the static approaches, we mostly consider the hand silhouette from a single image. Unlike static approaches, dynamic approach defines the gesture as a sequence of hand shapes, considering the aspects of hand motion. It helps to describe both the motion and the hand shape of the gesture. The first step is to capture the gesture using motion capture devices. Then comes the role of feature extraction to extract the most appropriate features from a whole set of features. A statistical representation method, FV and temporal representation method, TP are used to simplify complex calculations and improve performance. The classification is performed using an SVM classifier.

3.1 Capturing 3D Gestures

Motion capture devices which have sensors attached to a glove are the most reliable tools for capturing 3D hand gestures. But they have some drawbacks like cost, naturalness of hand, etc. The effective and inexpensive depth sensors, like the Microsoft Kinect, are popular today. Depth images also offer a great opportunity for the purpose. Shotton et al. [2] proposed a method to predict the 3D positions of 20 body joints, from depth images, called body skeleton. The most recent advancement is the device, such as Intel Real Sense or the Leap Motion Controller (LMC). This provides precise information of the full 3D skeleton corresponding to 22 joints. Since there are different possibilities for obtaining the precise skeletal data, it is an excellent choice to rely upon.

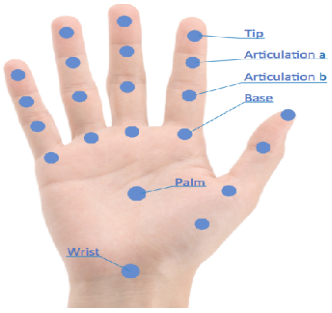


Figure 3.1: 22 Joints in hand

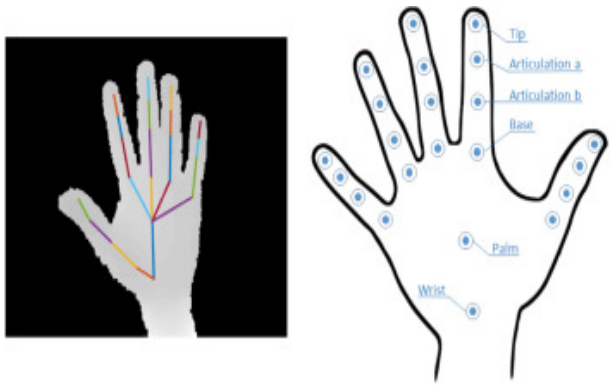


Figure 3.2: Depth and hand skeletal data returned by the Intel Real Sense camera

A dynamic gesture is a time series of hand skeletons. The hand shape and its motion is described along these series of input.
For each time frame t in the sequence, the position of Nj joints are represented using three coordinates as,

$$j_i(t) = [x_i(t)y_i(t)z_i(t)]$$

The hand skeleton at time frame t is represented by 3*Nj dimension row vector,where Nj is the number of joints in a single hand skeletal image,

$$s(t) = [\, x_1(t)y_1(t)z_1(t)...x_{Nj}(t)y_{Nj}(t)z_{Nj}(t)] \qquad (1)$$

Suppose the number of frames in a single sequence is Nf, then the whole sequence is represented using a matrix of size Nf*3Nj as,

$$M = \begin{bmatrix} s(1) \\ \vdots \\ (N_f) \end{bmatrix} \qquad (2)$$

In order to accurately represent the gesture, not only the hand shape is considered but also the direction of movement and rotation of hand are considered.

3.2 Feature Extraction

Motion Features

Some gestures, for instance swipes are expressed almost only by the way in which the hand moves in space. In such cases, we define a direction vector for each time frame t in the sequence.

$$\vec{d}_{dir}(t) = \frac{j_{palm}(t) - j_{palm}(t-c)}{\left\| j_{palm}(t) - j_{palm}(t-c) \right\|} \qquad (3)$$

where j palm is the position of palm joint

c is an experimentally chosen constant value

In equation (3), the direction vector is normalized by dividing it by its norm A set SD is defined for the direction vector for a sequence of Nf frames.

$$S_D = \left\{ \vec{d}_{dir}(t) \right\}_{[1 < t < N_f]} \qquad (4)$$

To take into consideration how the hand moves during the gesture, the rotation of the wrist is analyzed. For each time frame t, to get the rotational information a rotational vector from wrist node to palm node is computed,

$$\vec{d}_{rot}(t) = \frac{j_{palm}(t) - j_{wrist}(t)}{\left\| j_{palm}(t) - j_{wrist}(t) \right\|} \qquad (5)$$

A set SR is defined for the rotational vector of a sequence of Nf frames,

$$S_R = \left\{ \vec{d}_{rot}(t) \right\}_{[1 < t < N_f]} \qquad (6)$$

Hand Features

A descriptor based on joints is defined to represent shape of hand. It is called as Shape of Connected Joints (SoCJ). Here, a normalization phase is considered. Firstly, in order to take into account the differences of hand size between performers, we estimate the average size of each bone of the hand skeleton using all hands in the dataset. Firstly, using all hands in the dataset, an average of size of each bone is estimated. Secondly, the size is changed by their respective average size but keep the angles between bones unchanged.

It is important that the estimations are consistent with the translation and rotation transformation. Hence, consider a reference hand Hf with joints [0 0 0], called as root joint with its front facing the camera. The input image is translated and rotated so as to align it with the reference skeleton. This results in a new hand which is in reference to the base hand with joints [0 0 0]. The translation and the rotation with reference to the base hand is computed on for only the first hand of the sequence of the gesture. This calculation is applied to all other skeletons in the sequence of hand skeletons in a gesture sequence.

Suppose x is the coordinates of joints of hand skeleton, let there be 6 different joints as T=[x1 x2 x3 x4 x5 x6], the SoCJ is defined as the displacement from one joint to the previous joint,

$$\text{SoCJ(T)} = [x_2 - x_1 \ldots . x_6 - x_5] \qquad (7)$$

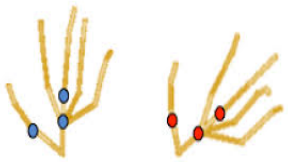


Figure 3.3: Before translation and rotation

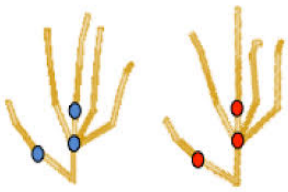


Figure 3.4: After translation and rotation

The use of Intel Real Sense camera provides with information of 22 joints of hand skeleton S. Theoretically, C(22,5)= 26334 different SoCJs for hand skeleton can be calculated, where C is a binomial coefficient function resulting in the set[2]

$$s_{socj} = \{ SoCJ(i) \}_{[1 < i < 26334]} \tag{8}$$

For a sequence of Nf frames, the set $S_{soc\,j}is$,

$$S_{socj} = \left\{ s_{socj}(t) \right\}_{[1 < t < N_f]} \tag{9}$$

3.3 Statistical Representation

The Fisher Vector (FV) coding method is relied upon for the purpose of statistical representation of data. The FV method was meant for large scale image classification when it was first introduced. Its superiority against the Bag-Of-Word (BOW) method has been analyzed in the image classification.[13]

In statistical classification, the Fisher kernel, named after Ronald Fisher, is a function that measures the similarity of two objects on the basis of sets of measurements for each object and a statistical model. In a classification procedure, the class for a new object (whose real class is unknown) can be estimated by minimising, across classes, an average of the Fisher kernel distance from the new object to each known member of the given class.[18]

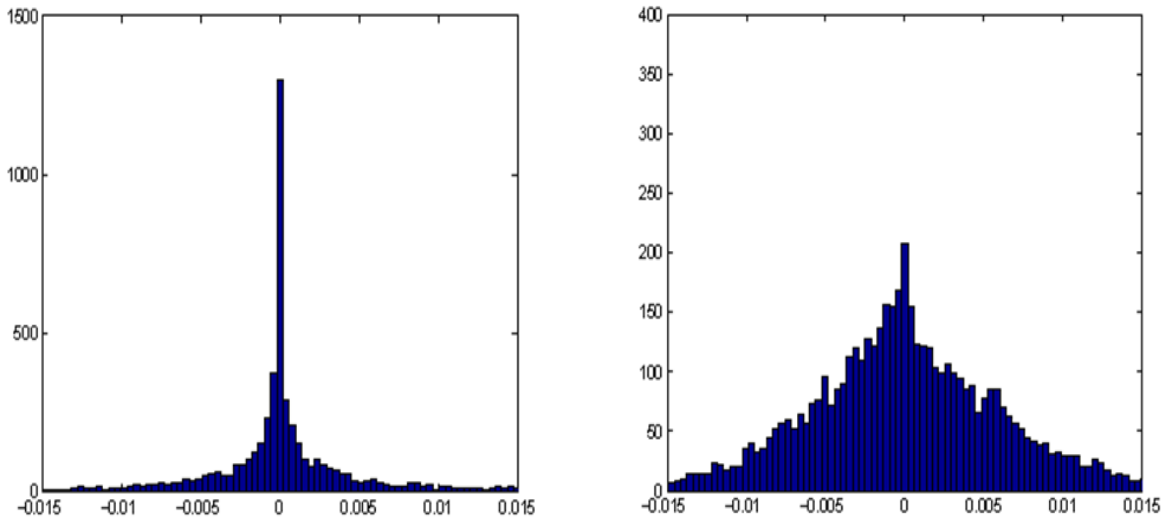


Figure 3.5: Fisher Vector: Beyond BoV

Fisher Vector encoding method characterizes a sample by its deviation from the generative model GMM.The deviation is measured by computing the gradient of the sample log-likelihood with respect to the model parameters (w,m,s).[16] Gaussian mixture models are a probabilistic model for representing normally distributed subpopulations within an overall population. Mixture models in general don't require knowing which subpopulation a data point belongs to, allowing the model to learn the subpopulations automatically. Since subpopulation assignment is not known, this constitutes a form of unsupervised learning.[8]

Gaussian mixture models are a probabilistic model for representing normally distributed subpopulations within an overall population. Mixture models in general don't require knowing which subpopulation a data point belongs to, allowing the model to learn the subpopulations automatically. Since subpopulation assignment is not known, this constitutes a form of unsupervised learning.[8]

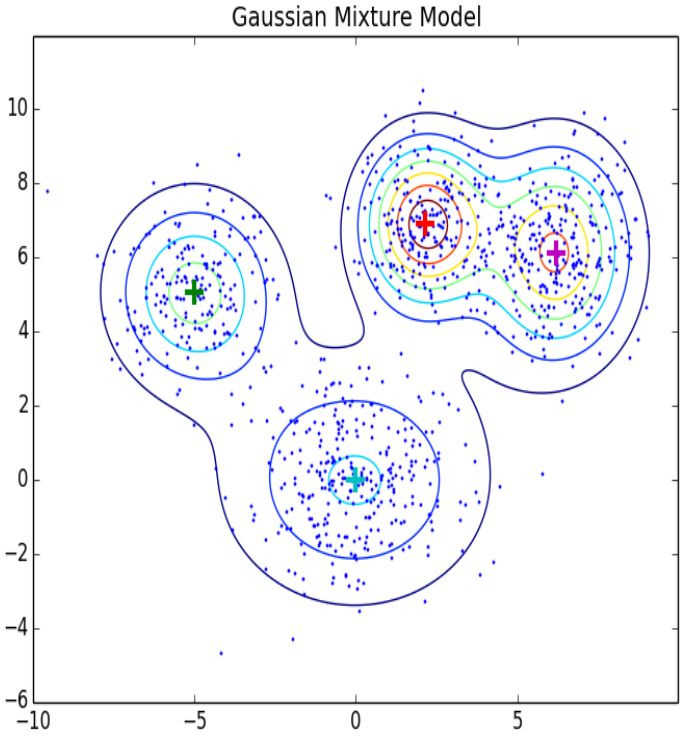


Figure 3.6: Gaussian mixture model

The first step is to train a K-component Gaussian Mixture Model. Then compute the FV. Then comes a normalization phase in order to eliminate sparseness of FV. While proceeding with FV it is found that the final size of an FV is 2*d*K where d is the size of feature of data and K represents the number of clusters in classification. This creates a drawback compared to the BOW, which has a size of just K, even when applied to a long descriptor. However, this drawback can be ignored. This is because in this case the value of K is relatively small.

The Fisher Vector representation has several applications over other methods. So, using this statistical representation method improves accuracy to some extend.

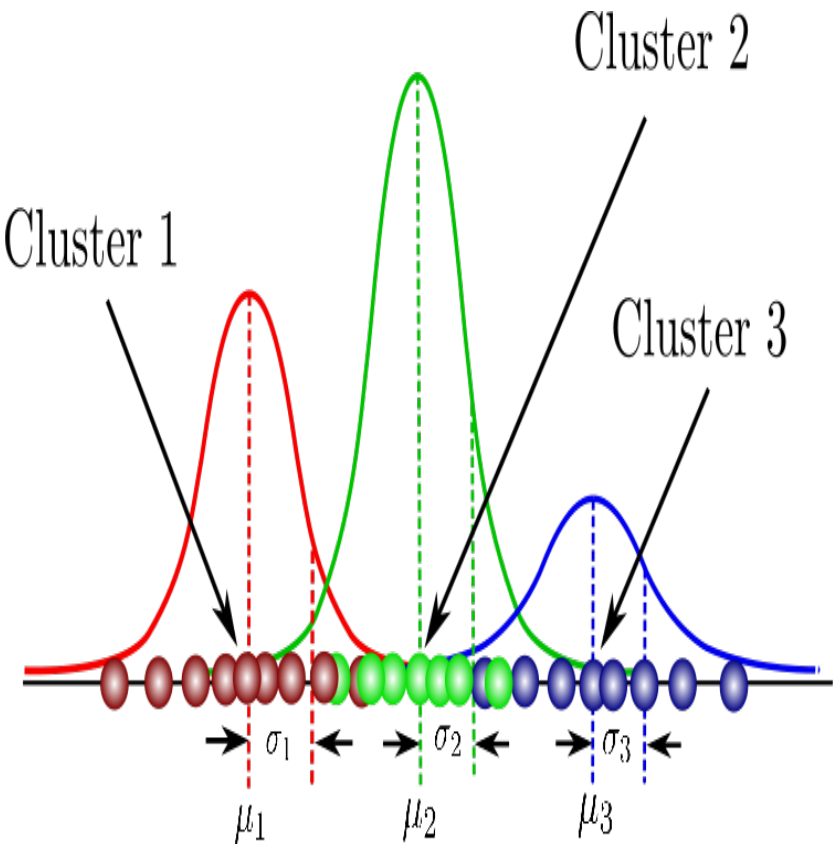


Figure 3.7: Gaussian mixture model

A standard approach to describe an image for classification and retrieval purposes is to extract a set of local patch descriptors, encode them into a high dimensional vector and pool them into an image-level signature. The most common patch encoding strategy consists in quantizing the local descriptors into a finite set of prototypical elements. This leads to the popular Bag-of-Visual words representation. In this work, we propose to use the Fisher Kernel framework as an alternative patch encoding strategy: we describe patches by their deviation from an "universal" generative Gaussian mixture model. This representation, which we call Fisher vector has many advantages: it is efficient to compute, it leads to excellent results even with efficient linear classifiers, and it can be compressed with a minimal loss of accuracy using product quantization.[13]

The sequence of hand gestures are represented by three sets of different features describing the direction of hand (SD), its rotation (SR) and its shape (Ssocj), after the feature extraction.

3.4 Temporal Representation

Now comes the stage to take into consideration the dynamic nature of hand. The use of a Temporal Pyramid (TP) representation, which is already employed in action and hand gesture recognition approaches, is to add a temporal cue.[3, 21]

The principle of the TP is to divide the sequence into a number of sub sequences. Suppose the pyramid has n levels. Each i th level of the pyramid will have i sub sequences. The three descriptors and their respective statistical representations are calculated for each sub sequence and the results are concatenated. More number of levels to the pyramid guarantees more temporal precision but increases computing time.

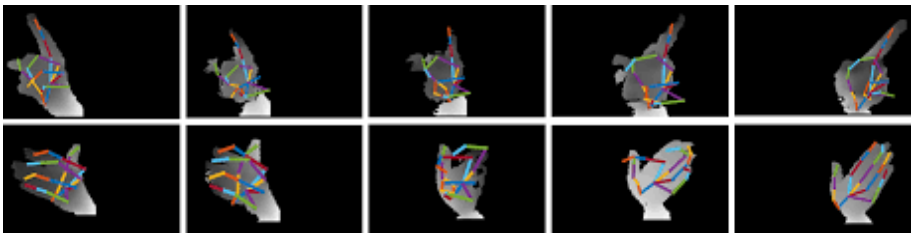


Figure 3.8: Swipe Right gesture performed (top) with one finger and (bottom) with the whole hand from the DHG-14/28 dataset.

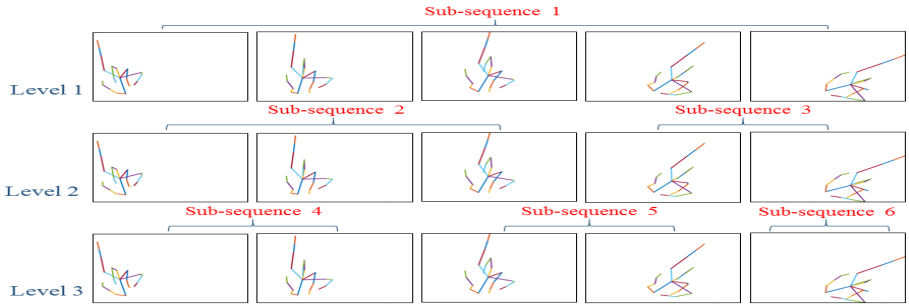


Figure 3.9: Temporal pyramid for swipe right gesture performed with one finger

A generic Temporal Pyramid Network (TPN) at the feature-level, which can be flexibly integrated into 2D or 3D backbone networks in a plug-and-play manner. Two essential components of TPN, the source of features and the fusion of features, form a feature hierarchy for the backbone so that it can capture action instances at various tempos. TPN also shows consistent improvements over other challenging baselines on several action recognition datasets. Specifically, when equipped with TPN, the 3D ResNet-50 with dense sampling obtains a 2per-cent gain on the validation set of Kinetics-400. A further analysis also reveals that TPN gains most

of its improvements on action classes that have large variances in their visual tempos, validating the effectiveness of TPN.[20]

3.5 Classifiaction

Classification is a predictive modelling problem where the model predicts the class of a random input. The classification process suggested here is through an SVM classifier. In machine learning, support-vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.[19] SVM with a linear kernel is used as the data high dimensional. A one-vs-rest strategy of SVM is used. That is, classifying a multi classification problem using a binary classification strategy. We group classes as one class and the second group with all left classes.

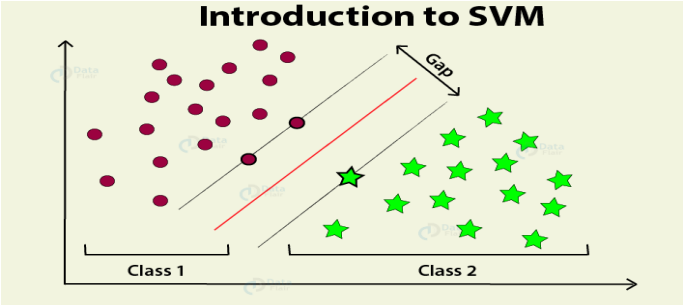


Figure 3.10: Basic SVM classification

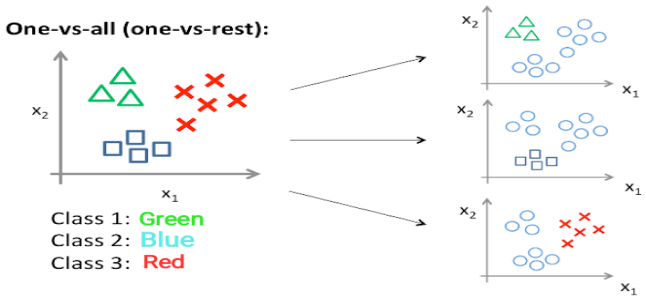


Figure 3.11: One vs. Rest SVM

3.6 Result

The The Intel Real Sense short range depth camera can be used to populate the dataset. Each frame gives coordinates of 22 hand joints in the camera space. To encode the descriptor, there is a need to fix the number of levels in temporal pyramid. For a hand skeleton with 22 joints, one can compute 26334 joints, But using all of them is not mandatory. The task is to choose the feature set as combination of the most relevant SoCJs. The first step is to intuitively select a SoCJ set anf then use a selection algorithm called Sequential Forward Floating Search (SFFS). The next task is to perform the tasks step by step.

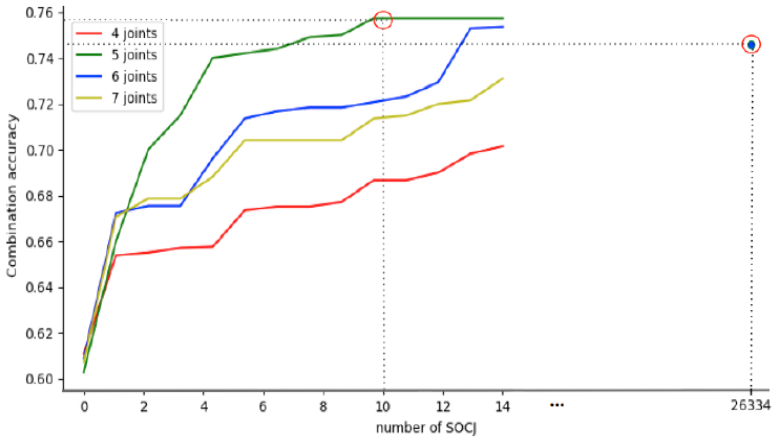


Figure 3.12: : SoCJ selection using SFFS algorithm on the fine gesture subset of the DHG dataset.

Each additional step in approach is to improve the accuracy of recognition. Here, the use of statistical representation method Fisher Vector and temporal representation method Temporal Pyramid improves the efficiency of the system. For Fisher Vector encoding, we map our descriptors into a K-component GMM with K equal to 8, 8 and 256 gaussians respectively for the direction, the rotation and the SoCJ features.[2]

The approximate accuracy percentage in the recognition system with and without FV and TP are given in the table below.

Features + SVM	76.89
Features + FV + SVM	79.76
Features + TP + FV + SVM	86.86

Figure 3.13: The approximate accuracy

CHAPTER 4
ADVANTAGES AND APPLICATIONS

4.1 Advantages

1. It takes into consideration the hand shapes, orientation and movement of hand.
2. Use of effective sensors provides precise information of hand skeleton.
3. Adding Fisher Vector representation and Temporal Pyramid representation increases the accuracy.

4.2 Applications

1. The sign language detection and recognition.
2. The hand gesture recognition for HCI.
3. The application in virtual environment control, especially in the field of gaming and entertainment.

CHAPTER 5
CONCLUSION

The approach for hand gesture detection and recognition using 3D dynamic skeletal data is analysed. The approach relies on the structure of hand topology to extract the effective descriptors from the gesture sequence. The method takes into consideration the hand shape, orientation and movement of hand. A dynamic gesture is represented as a series of hand skeletons. The hand skeletal data with information of 22 joints obtained from Intel Real Sense or LMC is used as the input. The appropriate features are extracted. The statistical representation method, Fisher Vector representation and the temporal representation method, temporal pyramid are used for the encoding of the descriptors. The final classification is done by the SVM classifier. The approach promises to enhance the performance of hand gesture detection and recognition system in terms of latency. The use of skeletal data of hand topology is an excellent choice towards the approach for hand gesture recognition system.

REFERENCES

[1] Quentin De Smedt, Hazem Wannous, and Jean-Philippe Vandeborre. Skeleton-based dynamic hand gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2016.

[2] Quentin De Smedt, Hazem Wannous, and Jean-Philippe Vandeborre. Heterogeneous hand gesture recognition using 3d dynamic skeletal data. *Computer Vision and Image Understanding*, 181:60–72, 2019.

[3] Jorge García, Niki Martinel, Gian Luca Foresti, Alfredo Gardel, and Christian Micheloni. Person orientation and feature distances boost re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 4618–4623. IEEE, 2014.

[4] Dariu M Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

[5] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3593–3601, 2016.

[6] Haiying Guan, Rogério Schmidt Feris, and Matthew Turk. The isometric self-organizing map for 3d hand pose estimation. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 263–268. IEEE, 2006.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Andrei Dobre John McGonagle, Geoff Pilling and 5 others contributed. Gaussian mixture model: https://brilliant.org/wiki/gaussian-mixture-model/: :text=gaussian

[9] Adam Kendon. The biological foundations of gestures: Motor and semiotic aspects, 1986.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[11] 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.

[12] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. Action and event recognition with fisher vectors on a compact feature set. In *Proceedings of the IEEE international conference on computer vision*, pages 1817–1824, 2013.

[13] Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classi-

fication with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[15] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):1–10, 2014.

[16] Gonzalo Vaca-Castano. Fisher vector encoding: http://www.cs.ucf.edu/courses/cap6412/spr2014/papers/fisher-vector-encoding.pdf.

[17] Christian Vogler and Dimitris Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *Computer Vision and Image Understanding*, 81(3):358–384, 2001.

[18] Wikipedia. Fisher vector: https://en.wikipedia.org/wiki/fisher-kernel.

[19] Wikipedia. Support vector machine: https://en.wikipedia.org/wiki/support-vector-machine.

[20] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 591–600, 2020.

[21] Chenyang Zhang, Xiaodong Yang, and YingLi Tian. Histogram of 3d facets: A characteristic descriptor for hand gesture recognition. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.