Multimodal Information Fusion for Human Action Recognition

by

Nour Eldin Elmadany

Master of Science, Arab Academy for Science and Technology, 2012

Bachelor of Science, Arab Academy for Science and Technology, 2008

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2019

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A DISSERTATION

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Multimodal Information Fusion for Human Action Recognition

Doctor of Philosophy, 2019

Nour Eldin Elmadany

Electrical and Computer Engineering

Ryerson University

Abstract

This thesis presents three frameworks of human action recognition to facilitate better recognition performance. The first framework fuses handcrafted features from four different modalities including RGB, depth, skeleton, and accelerometer data. In addition, a new descriptor for skeleton data is proposed that provides a discriminative representation for the poses of an action. Since the goal of the first framework is to find a more discriminative subspace, a generalized fusion technique Multimodal Hybrid Centroid Canonical Correlation Analysis (MHCCCA) is proposed for two or more sets of features or modalities. The second framework fuses handcrafted and deep learning features from three modalities including RGB, depth, and skeleton. In this framework a new depth representation is introduced that extracts the final representation using Deep ConvNet. The proposed fusion technique forms the backbone of the framework: Multiset Globality Locality Preserving Canonical Correlation Analysis (MGLPCCA) for two or more sets of features or modalities. MGLPCCA aims to preserve the local and global structures of data while maximizing the correlation among different modalities or sets. The third framework uses the deep learning techniques to improve the long term temporal modelling through two proposed techniques: Temporal Relational Network (TRN) and Temporal Second Order Pooling Based Network (T-SOPN). Additionally, Global-Local Network (GLN) and Fuse-Inception Network (FIN) are proposed to encourage the network to learn complementary information about the action and scene itself. Qualitative and quantitative experiments are conducted on nine different datasets demonstrating the effectiveness of the proposed framework over state-of-the-art methods.

Acknowledgements

First and foremost, I would like to thank Canada; I will always be indebted to this country, my second home country. Represented by Ryerson University, Canada provided me unlimited resources of knowledge in order to become a better educated person.

I would like to thank my dear supervisor, Prof. Ling Guan, for everything he has done for me, for his support when I faltered, for his unconditioned financial support, for bearing my unprofessionalism many times, for his knowledge and time, for the doors he opened for me, for the several last-minute paper reviews, the personal level in his friendly discussions and for being patient while reviewing my poorly written manuscripts. In fact, any success I have achieved or will achieve throughout my life will bear the signature of Prof. Ling Guan. Moreover, I would like to express my sincere thanks to Dr. Yifeng He for his support, valuable suggestions, inspirational ideas throughout my studies, and his careful editing of manuscripts. Also, I would like to extend my sincere thanks to Prof. Yan Lu from Microsoft Research Asia (MSRA); for being a great mentor while I was interning at MSRA and for the great discussions I had with him. I would like to thank my research lab-mate, Gareth Higgins, for reviewing my manuscripts, and his wonderful calming conversations. I thank Lie Gao and Liang Chengwu for their scientific brainstorming conversations we had and technical advices.

I would like to thank all my lab-mates (current and alumni) and I wish them all flourishing careers. I would like to thank Prof. Mehmet Zeytinoglu for his help and advice.

My sincere thanks to the examiners of my thesis, Prof. Fahkri Karray, Prof. Dimitri Androutsos, Prof. Javad Alirezaie, Prof. Kosta Derpanis, and Dr. Foivos Xanthos, for their thorough review and insightful comments.

I thank my friends Amr Rizk, Khaled Elzafraany, and Tamer Badran for their long distance healing calls and support.

Ink always falls short of expressing my deep love and gratitude to my parents and brother for their unwavering love, long distance healing calls and wholehearted support. I thank my uncles and aunts for being supportive through my studies. Especially, my uncle Mohamed who acted as a role model and career supporter. None of this would have been possible without them.

Contents

List of Tables

List of Figures

Acronyms

BoA Bag of Angles. xiv, 4, 25, 27, 32–35, 37–39, 43, 60–62

BoF Bag of Features. 9

C3D 3D Convolutional Networks. 11

CCA Canonical Correlation Analysis. xi–xiii, xv, 2, 16–18, 33, 34, 37, 40–43, 48, 59–65, 90, 92

CCCA Centroid Correlation Analysis. xi, xii, 16, 18, 34, 37, 41, 42

CRF Conditional Random Fields. xv, 5, 68–70, 90, 92

DMM Depth Motion Map. 11, 12, 14, 26, 27

FIN Fuse-Inception Network. iv, xvi, 6, 67, 69, 73, 78, 81, 88, 91

GLN Global-Local Network. iv, 5, 67, 68, 73, 80, 83, 88, 91

GLPCCA Globality Locality Preserving Canonical Correlation Analysis. viii, 5, 44, 48, 49

GLPP Globality Locality Preserving Projections. 45, 46

GMM Gaussian of Mixture Models. 9

GPP Globality Preserving Projections. 45

HCCCA Hybrid Centroid Canonical Correlation Analysis. 4, 17–19, 21, 22, 24

HCRF Hidden Conditional Random Field. 92

HOD Histogram of Oriented Displacement. 13

HOF Histogram of Oriented Optical flows. 9

HOG Histogram of Oriented Gradients. 9, 12, 27

HON4D Histogram of Oriented 4D Normals. 12, 58

HP-DMM Hierarchical Pyramid DMM. xiv, 25, 26, 28, 33–35, 39, 42, 53

HP-DMM-CNN Hierarchical Pyramid DMM Deep Convolutional Neural Network. 53, 55, 58, 60–65, 90

JDM Joint Distance Maps. 13

JTM Joint Trajectory Maps. 13

KCCA Kernel Canonical Correlation Analysis. 2

KLT Kanade-Lucas-Tomasi. 8

LASC Locality-constrained Affine Subspace Coding. 12

LBP Local Binary Pattern. 12

LLC Locality-constrained Linear Coding. 14

LPCCA Locality Preserving Canonical Correlation Analysis. 60, 61, 63, 64

LPP Locality Preserving Projections. 45

LSTM Long Term Short Memory. 3, 11, 13

MBH Motion Boundary Histogram. 9

MCCA Multiset Canonical Correlation Analysis. xi, xiv, 2, 22, 32, 33, 37, 40, 42, 43, 50, 60, 90

MCCCA Multimodal Centroid Canonical Correlation Analysis. xi, 37, 40, 42

MEI Motion Energy Image. 8, 25, 27

MGLPCCA Multiset Globality Locality Preserving Canonical Correlation Analysis. iii, ix, xii, xiii, xv, 5, 6, 45, 50, 52, 55, 59–65, 90, 92

MHCCCA Multimodal Hybrid Centroid Canonical Correlation Analysis. iii, xi, xii, xiv, 4, 6, 15–17, 22, 24, 32–34, 36–38, 40–43, 89, 92

MHI Motion History Image. 8, 10, 12, 25, 27, 35, 37, 38, 41, 42

MLPCCA Multiset Locality Preserving Canonical Correlation Analysis. 60, 63

PCA Principal Component Analysis. 60

ReLU Rectified Linear Unit. 54

ROP Random Occupancy Pattern. 12

SEN Squeeze and Excitation Networks. 11

SF Statistical Features. xiv, 25, 29, 35–39, 41, 42

SGD Stochastic Gradient Descent. 80

SIFT Scale Invariant Feature Transform. 9

SMSO Statistically Motivated Second Order. 71, 72

STOP Space Time Occupancy Patterns. 12

t-SNE t-distributed Stochastic Neighbor Embedding. 32, 59, 65

T-SOPN Temporal Second Order Pooling Based Network. iv, 5, 67–69, 71, 80, 83, 88

TRN Temporal Relational Network. iv, xv, 5, 66, 68–70, 80, 83, 88, 90

TSN Temporal Segment Network. 10

TSOP Temporal Second Order Pooling. xv, 71, 72, 82

VLAD Vector of Locally Aggregated Descriptors. 9

Chapter 1

Introduction

1.1 Background

Human action recognition has witnessed major developments recently due to its importance in a variety of applications across different fields including: health care [1], security [2], and human-robot interaction [3]. However, human action recognition is a challenging research topic due to the presence of many problems such as: occlusion, varying light conditions, the sensor noise, and speed variations while performing actions. Researchers have mainly focused on human activity analysis from RGB cameras, including methods based on human silhouettes [4] [5], spatial-temporal shapes [6] and local based descriptors [7]. Though various methods have been proposed for RGB videos, there are still limitations including sensitivity to different illuminations.

Researchers have invested substantial effort into developing 3D action recognition. The last decade has witnessed the development of inexpensive depth cameras such as Kinect. Kinect cameras have a distinct advantage over traditional RGB cameras because they capture 3D human motion in space using depth videos and skeleton data. Additionally, several methods were proposed based on temporal motion history [8], or local motions in videos using local interest points [9]. Many researchers have proposed techniques based on skeleton data to capture the differences between joints [10] or describe the trajectory of an action [11]. However, depth videos have the problems with noise and occlusion. As a result, other researchers have been investigating human action recognition using inertial sensors that can provide accurate acceleration data and angular velocity.

Each sensor captures a different aspect of the subject performing the action, for example RGB cameras capture the scene and texture information regarding the subject, while depth cameras capture the relations among body parts. A possible method for solving these challenges in human action recognition is to find a way of incorporating different modalities representing human action more effectively. Several researchers proposed to simultaneously utilize different modalities like depth and inertial sensors to improve the recognition accuracy [12].

However, simple serial or parallel feature fusion can not identify the intrinsic relations among different modalities. Some researchers proposed techniques which learn the intrinsic relations among different sets of data. Hardoon *et al.* proposed Canonical Correlation Analysis (CCA) to maximize the correlation between two different sets of data [13]. Furthermore, Kernel Canonical Correlation Analysis (KCCA) [13] revealed the nonlinear relationship between two sets of data. Nevertheless, methods such as CCA and KCCA are restricted to finding the relationships between only two sets. Multiset Canonical Correlation Analysis (MCCA) [14] was proposed to reveal the relationship among multiple sets of data or modalities[1]. However, because they do not incorporate the knowledge of class labels during training, the aforementioned methods fail to find a discriminative common space which can differentiate similar actions. For example, CCA, KCCA, and MCCA have difficulty in separating the actions between sitting down and standing up, which are two reversely-ordered actions with a small between-class

[1] Throughout the dissertation, set is used interchangeably with modality.

distance in the common space.

Recently, researchers have began to apply deep learning in action recognition. Researchers introduced 2D Deep ConvNets [15] in action recognition from RGB videos. However 2D Deep ConvNets can not capture the long term temporal information. Others introduced 3D Deep ConvNets [16] which require a large number of frames to achieve optimum performance. Moreover, others explored the cooperation between 2D Deep ConvNets [17] and Long Term Short Memory (LSTM) [18] for action recognition.

1.2 Purpose and Scope of This Dissertation

The work of this dissertation focuses on improving recognition accuracy, with a goal of building long term effective frameworks for action recognition. Specifically, more effective action recognition based on the fusion of different features; in particular depth, skeleton, RGB, and accelerometer data is considered. Moreover, the impact of deep learning techniques for feature representation and learning is explored. Additionally, fusion techniques are utilized to further improve the action representations in deep learning techniques. Overall, the thesis aims to improve the human action recognition performance. The presented frameworks are based on the fusion of different modalities or sets. They represent a seamless migration journey in human action recognition from handcrafted features techniques for small scale datasets to deep learning techniques for large scale datasets. The first framework depends only on handcrafted features extracted from RGBD cameras; which were trending at the time. The second framework represents an intermediate stage where deep learning is merged with handcrafted. The third framework demonstrates the dominance of deep learning techniques in action recognition in large datasets. To demonstrate the effectiveness of the proposed frameworks they were tested on several different datasets; with different groups of modalities.

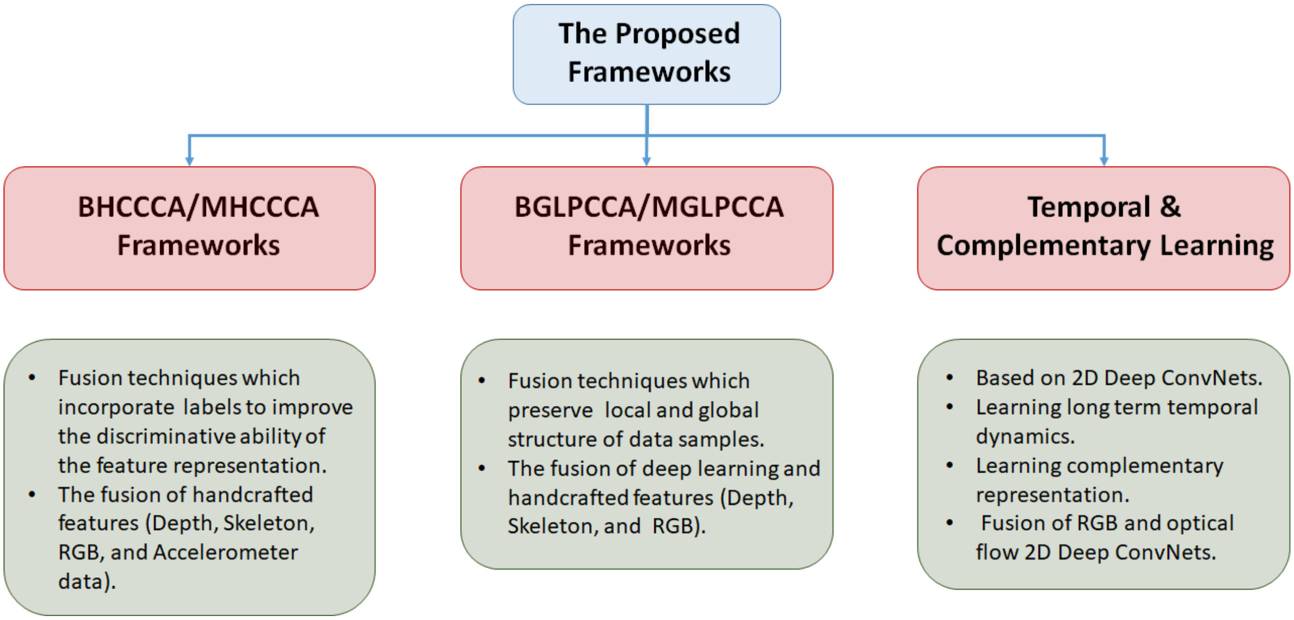

The Proposed Frameworks
BHCCCA/MHCCCA Frameworks
BGLPCCA/MGLPCCA Frameworks
Temporal & Complementary Learning
• Fusion techniques which incorporate labels to improve the discriminative ability of the feature representation.
• The fusion of handcrafted features (Depth, Skeleton, RGB, and Accelerometer data).
• Fusion techniques which preserve local and global structure of data samples.
• The fusion of deep learning and handcrafted features (Depth, Skeleton, and RGB).
• Based on 2D Deep ConvNets.
• Learning long term temporal dynamics.
• Learning complementary representation.
• Fusion of RGB and optical flow 2D Deep ConvNets.

Figure 1.1: Dissertation contributions diagram.

1.3 Contributions

The contributions of this work are summarized in Figure 1.1. The key areas of benefit are as follows:

- **A novel human action recognition framework that fuses different modalities (RGB, depth, skeleton, and accelerometer data) is proposed**. The fusion of the features is based on the proposed fusion technique, named Hybrid Centroid Canonical Correlation Analysis (HCCCA), for learning discriminative common space between two different modalities. HCCCA uses the class labels of the training dataset to find the discriminative common feature space. HCCCA is then generalized to Multimodal Hybrid Centroid Canonical Correlation Analysis (MHCCCA) to deal with multiple feature sets. This framework explores the fusion of handcrafted features in performance boosting. Additionally, a Bag of Angles (BoA) is proposed to represent skeletons, aiming to capture the key distinct human poses representing the characteristics of an action from skeleton sequences.

- **A human action recognition framework which combines heterogeneous features from different modalities or sets (RGB, depth, and skeleton) using both deep learning and handcrafted features is presented**. This framework highlights the effectiveness of fusion by using the proposed Globality Locality Preserving Canonical Correlation Analysis (GLPCCA) which learns the common subspace between two different sets, preserving both the local structure (the geometric relations among the data samples in the same class) and global structure (the geometric relations among different classes). Moreover, Multiset Globality Locality Preserving Canonical Correlation Analysis (MGLPCCA) is presented for three or more sets or modalities. The framework exploits 2D Deep ConvNets as a feature extractor for both RGB and depth videos.

- **A new deep learning framework for RGB videos based on two stream networks is introduced**. The presented framework improves both the long term temporal dynamics and leverages complementary information to improve the discriminative ability of the learned representation. To capture the long term temporal dynamics, two temporal learning networks are presented for better long term temporal modelling. First, Temporal Relational Network (TRN) which combines 2D Deep ConvNets and Conditional Random Fields (CRF) in an end-to-end training framework is presented. Second, Temporal Second Order Pooling Based Network (T-SOPN) is proposed to model the temporal structure of the video. T-SOPN models the long term temporal dynamic cues through pooling. Additionally, to improve the overall recognition performance of the video, two effective methods are proposed based on the idea of complementary learning. Global-Local Network (GLN) is proposed to combine both global information and the local discriminative features corroborating the discriminative representation. In addition, Fuse-Inception Network (FIN) is presented to learn the feature representation through fusion and complementary learning.

1.4 Overview of This Dissertation

An overview of the organization of the remainder of the dissertation follows. Chapter 2 reviews existing work related to action recognition. Chapter 3 presents the proposed action recognition based on MHCCCA, and compares the framework against other similar methods. Chapter 4 introduces our human action recognition framework which fuses deep learning and handcrafted features using the proposed MGLPCCA, and tests the effectiveness of the framework. Chapter 5 introduces the temporal and complementary learning techniques that are based on deep learning and explores the effectiveness of the proposed techniques against state-of-the-art techniques. Finally, Chapter 6 encapsulates the conclusions and discusses possible future work.

Chapter 2

Related Work

The most relevant, and prominent techniques are reviewed in order to set the stage for the investigation and development of human action recognition, and to acknowledge the contributions in this field. Over the last three decades, researchers have proposed a variety of human action recognition techniques. In this dissertation, the existing works are divided according to the input modalities they use, into four main categories: RGB, depth, skeleton, and hybrid based techniques.

2.1 RGB Based Techniques

Researchers have presented a diverse set of techniques for human action recognition, which can be classified into two main groups, handcrafted based methods and deep learning based methods.

2.1.1 Handcrafted Based Methods

Shape and silhouette features are considered to be the first proposed human action representations. Bobick *et al.* [4] introduced the idea of summarizing human actions over time through temporal templates. They extracted human shape masks from a video sequence and cumulatively summed the difference between each consecutive pair of frames to construct Motion History Image (MHI) and Motion Energy Image (MEI), representing the presence and recency of the action. Blank *et al.* [6] extended 2D silhouette to the 3D silhouette representing the action over time as a space-time volume. However, these methods are highly dependent on a precise human body segmentation.

Another approach is to focus on local feature methods which do not require any pre-knowledge regarding the human body. These are based on local feature detection and spatio-temporal descriptors to represent the detected features. Laptev *et al.* [19] extended Harris detector [20] to Harris3D interest point detector. The Harris3D detector finds the local positive spatio-temporal maxima by computing a spatio-temporal second moment matrix at each point, and searches for active video regions by finding the largest eigenvalues of this matrix. Dollar *et al.* [21] proposed behaviour recognition using Gabor filters, which gives more dense results than Harris3D detector, by applying a set of spatial Gaussian kernels and temporal Gabor filters. Willems *et al.* [22] proposed to use Hessian3D as a local spatio-temporal detector in action recognition, which is an extension to the Hessian saliency measure of blob detection in images [23]. The Hessian3D detector computes the Hessian matrix at each interest point, then calculates the determinant of the Hessian matrix for point localization. Wang *et al.* [24] proposed to densely sample interest points at regular positions and scales in time. Others proposed detection of the robust interest points by detecting the trajectories of the interest points. Matikainen *et al.* [25] tracked features using the Kanade-Lucas-Tomasi (KLT) [26] tracking algorithm, then clustered the trajectories and assigned to the library of trajectories. Sun *et al.* [27] proposed to extract trajectories based on pairwise matching over two consecutive frames. Wang *et al.* [28] proposed to densely sample interest points and track them using an optical flow field. Others studied local feature descriptors for videos, for example, Laptev *et al.* [7] proposed several descriptors based on motion representation. In [29], Laptev *et al.* proposed local feature descriptors including Histogram of Oriented Gradients (HOG) and Histogram of Oriented Optical flows (HOF). The former encodes the visual appearance and shape information and the latter encodes motion information. Scovanner *et al.* extended Scale Invariant Feature Transform (SIFT) to 3DSIFT [30]. 3DSIFT is based on the spatio-temporal grid concept and spatio-temporal gradients. Similarly, Klaser *et al.* extended the HOG [31] to HOG3D. Additionally, Wang *et al.* [32] proposed Motion Boundary Histogram (MBH) to spatio-temporal information in videos. Other researchers focused on local features encoding techniques for video representation. Niebles *et al.* [33] applied Bag of Features (BoF) on action recognition. In [33], the global representation was computed for the local features describing the sub-actions in the video. Then, a visual vocabulary over all extracted local features was created. Finally, the video was represented as a histogram of the quantified local features. Perronnin *et al.* [34] proposed Fisher vector encoding to encode the difference between the features and visual words. A visual vocabulary was created by clustering local features extracted from the video using Gaussian of Mixture Models (GMM) then the first and second moments of the differences between the local features and visual vocabulary were computed. Similarly, in [35], Jegou *et al.* introduced Vector of Locally Aggregated Descriptors (VLAD). It accumulates the residual of each local feature with respect to its assigned visual word. For each cluster, it stores the accumulated sum of the differences of the descriptors assigned to the cluster and the centroid of the cluster as well.

2.1.2 Deep Learning Based Methods

Over the past five years, researchers invested significant effort in video based action recognition using deep learning. Deep learning based methods can be roughly categorized according to their architecture into two main categories: 2D Deep ConvNets and

3D Deep ConvNets approaches.

2D Deep ConvNets approaches

Karpathy *et al.* [36] introduced a new dataset containing 1 million YouTube videos and empirically studied 2D Deep ConvNets. However, their work was to some extent limited in the temporal modelling. Simonyan *et al.* [37] introduced a two stream network with two 2D Deep ConvNets (Spatial for RGB) and (Temporal for optical flow). Their objective was to capture the complementary information on appearance and motion from RGB images and from optical flow, respectively. In [38], Feichtenhofer *et al.* studied different possibilities for fusing two stream 2D Deep ConvNets (Spatial for RGB) and (Temporal for optical flow), and found that the fusion of learned features from the last convolutional layer boosted performance. In order to improve the temporal modelling, in [39] and [40] researchers introduced spatio-temporal ResNet which injects residual connections from spatial to temporal and vice versa. Wang *et al.* [41] introduced a trajectory pooled deep convolutional descriptor which combined the merits of handcrafted and deep learning features. They utilized 2D Deep ConvNets to learn discriminative convolutional feature maps, and applied trajectory constrained pooling to aggregate these features into effective descriptors.

Zhang *et al.* [42] introduced motion vector based action recognition, which models the temporal information using motion vectors instead of optical flow. In [15], the authors introduced Temporal Segment Network (TSN) and set guidelines for good practices in training 2D Deep ConvNets for better and low cost performance. In [43], Diba *et al.* introduced the idea of deep temporal linear encoding, where bilinear pooling [44] was adopted to pool features across time. The introduced architecture encodes the appearance and motion through the time for a more compact representation. Sun *et al.* [45] introduced optical guided networks which can implicitly compute the optical flow. Bilen *et al.* [46] introduced a network which computes MHI implicitly using rank pooling. This provided 2D Deep ConvNet with the ability to capture evolution of the action itself over time. In [47], Zhou *et al.* modelled the temporal information through sum pooling the feature representations across time and at different scales. Donahue *et al.* [48] introduced the idea of modelling temporal cues by incorporating LSTM and represented video frames using 2D Deep ConvNets.

3D Deep ConvNets approaches

Tran *et al.* proposed 3D Convolutional Networks (C3D) to model the temporal dynamics and learn spatio-temporal features [49] through 3D convolutional and pooling layers. Varol *et al.* [50] studied the effectiveness of capturing the long term temporal information on the accuracy performance. Carreira *et al.* [16] extended Inception V1 [51] and introduced two stream Inflated 3D ConvNet (I3D) enabling spatio-temporal feature learning. In [52], Hara *et al.* introduced the 3D versions of ResNet [53], Wide ResNet [54], ResNext [55], and DenseNet [56]. Wang *et al.* introduced Non-Local Networks which capture long range temporal dependencies [57]. In [58], the authors adopted graph networks to learn the relationships among the features learned by I3D. Diba *et al.* [59] introduced a new layer called the temporal transition layer for 3D DenseNet to exploit temporal cues. Diba *et al.* extended Squeeze and Excitation Networks (SEN) [60] to a 3D version and applied it to action recognition. In [61], the authors introduced the Multi-Fiber Network which is a computationally efficient 3D ConvNet architecture. They sliced the complex architecture of 3D Deep ConvNets into lightweight networks called fibers.

2.2 Depth Based Techniques

In the depth based techniques, depth maps are used for activity recognition. In [8], Yang *et al.* proposed Depth Motion Map (DMM) as a human action representation which is the depth version of MHI. In DMM, depth maps are projected onto three orthogonal planes and the differences between successive frames are accumulated generating three DMMs. HOG was utilized to represent the three DMMs. However, DMM loses the temporal order information. Wang *et al.* [9] extracted semi-local Random Occupancy Pattern (ROP) and encoded the temporal information using sparse coding. In addition, Vieira *et al.* [62] proposed Space Time Occupancy Patterns (STOP) in which the space and time axes are divided into multiple segments for each sequence. The authors aimed to preserve spatial and temporal contextual information between space time cells. In [63], Chengwu *et al.* divided the sequence into energy equal sub-sequences. DMM was calculated for each sub-sequence. Then, each DMM is represented using gradient local auto-correlation features and Locality-constrained Affine Subspace Coding (LASC). In [64], Chen *et al.* proposed using Local Binary Pattern (LBP) instead of HOG as a descriptor. The fusion of features from different DMMs was utilized. In [65], Oreifej *et al.* described the depth sequence using the histogram capturing the distribution of the normal surface orientation in the 4D space of time, depth, and spatial coordinates called Histogram of Oriented 4D Normals (HON4D). However, the HON4D is highly computationally intensive. In [66], Tran *et al.* presented and proposed to incorporate optical flow in depth. Deep ConvNets were applied in human action recognition from depth videos. Wang *et al.* [17] proposed to train 2D Deep ConvNets on DMMs. In [67], the authors adopted scene flow as an action representation and trained 2D Deep ConvNets for classification.

2.3 Skeleton Based Techniques

In skeleton based techniques, action recognition is based on skeleton data only. Xia *et al.* [68] represented the posture using a histogram of 3D joint locations and the Hidden Markov Model was adopted for classification. They claimed that their proposed framework is view invariant. In [69], Ofli *et al.* proposed to select the most informative set of joints that can describe a certain action using highly interpretable measures including variance and mean of joint angle trajectories. Chaaraoui *et al.* [70] used an evolutionary algorithm to find the most informative joints considering the topological structure of the skeleton. In [10], Yang *et al.* proposed a set of features based on the differences of joints. They captured the static and dynamic postures through calculating the joint differences. The authors adopted the Naive Bayes Nearest Neighbor classifier. Others aimed to describe the trajectory of the joints over the video sequence. In [71], Hussein *et al.* represented the trajectory of the joints over the sequence using a covariance descriptor. The authors encoded the relationship between the joint location and time by computing multiple covariance matrices over sub-sequences of the skeleton trajectories. Similarly, Gowayyed *et al.* [11] proposed to model the trajectory of joints over the sequence using Histogram of Oriented Displacement (HOD). In HOD, each displacement in the trajectory votes with its magnitude in a histogram of oriented angles. More recently, LSTM was used for temporal modelling of the skeleton sequence. Liu *et al.* [72] proposed to use spatio-temporal LSTM for 3D action recognition. Du *et al.* [73] introduced a hierarchical recurrent neural network which represents the trajectories of skeleton joints where each skeleton is divided into five parts corresponding to the body parts. Wang *et al.* [74] proposed to represent spatio-temporal sequences as Joint Trajectory Maps (JTM) and adopted 2D Deep ConvNets for classification. Li *et al.* [75] introduced Joint Distance Maps (JDM) to represent the spatio-temporal information and adopted 2D Deep ConvNets for classification. Yan *et al.* [76] modelled the dynamics of skeleton through a spatio-temporal graph convolutional network which learns spatial and temporal patterns of actions.

2.4 Hybrid Based Techniques

In hybrid-based techniques, human action is recognized based on a mixture of different modalities, such as RGB, depth, skeleton, and accelerometer data. Zhu *et al.* [77] adopted Random Forests to fuse joint differences and spatio-temporal interest points including: Harris3D, HOG3D, HOG, and Hessian. In [12], Chen *et al.* introduced an action recognition framework that fuses accelerometer data and depth maps. They represented depth maps using DMM and accelerometer data using statistical features. The computed features were fused both at the feature and decision levels. In [78], Wang *et al.* proposed a framework based on local patterns of depth sequences and relative positions of the skeleton joints, to minimize the intra-class distances. Shahroudy *et al.* [79] fused RGB and skeleton data using unsupervised structured sparsity feature fusion which is based on Locality-constrained Linear Coding (LLC). Ofli *et al.* [80] presented a dataset which contains different actions captured by multiple modalities including depth, RGB cameras, accelerometers, and microphones then used multiple kernel learning to fuse the different modalities. Shahroudy *et al.* [81] introduced a deep autoencoder based on shared-specific feature factorization to fuse RGB and depth modalities.

Chapter 3

Action Recognition via MHCCCA

3.1 Introduction

The aim of this dissertation is to improve the accuracy performance of human action recognition using information obtained from multiple sources. Such improvement can be achieved by fusing different available modalities [82]. In this chapter, a new framework for human action recognition based on fusion of handcrafted features extracted from different modalities is introduced for small scale datasets. In this framework, four modalities specifically RGB, depth, skeleton, and accelerometer data are fused using the proposed novel fusion technique, Multimodal Hybrid Centroid Canonical Correlation Analysis (MHCCCA). The proposed fusion technique aims to learn the discriminative and informative shared space, by considering the correlation among different classes across two or more modalities. The proposed human action recognition framework as shown in Figure 3.1 consists of three stages: feature extraction, fusion, and finally classification. The proposed fusion technique MHCCCA forms the core of the framework, the chapter focuses first on the proposed fusion followed by the feature extraction and classification.

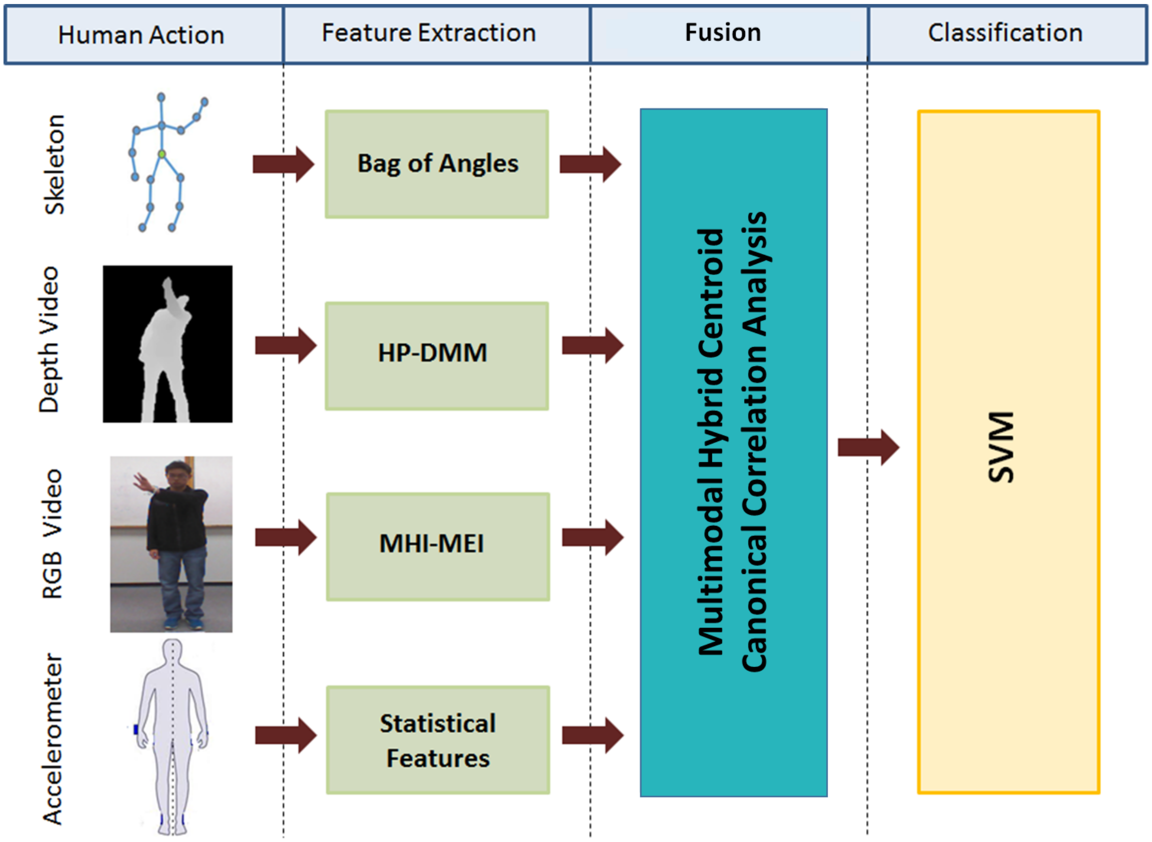


Figure 3.1: The proposed human action recognition framework using MHCCCA considers four modalities: RGB video, depth video, skeleton, and accelerometer data.

3.2 Hybrid Centroid Canonical Correlation Analysis

The idea of fusion is to find a common shared space representation of features extracted from different modalities or sets. The major challenge is to find the complementary relationship among the heterogeneous modalities or sets. In this section, the MHCCCA is introduced for feature fusion. The power of MHCCCA is its ability to learn the discriminative common low dimensional representation of the heterogeneous modalities or sets. For the sake of completeness, we first briefly describe CCA and Centroid

Correlation Analysis (CCCA), and then present the proposed HCCCA and MHCCCA.

3.2.1 Review on Canonical Correlation Analysis

Canonical correlation analysis (CCA) aims to maximize the correlation between two different sets [83] [13]. Consider two sets of n zero-mean centred samples $X \in \Re^{p \times n}$ and $Y \in \Re^{q \times n}$, p and q are the dimensions of feature samples in X and Y, respectively. CCA aims to learn the projection basis functions that maximize the correlation between the projected samples. CCA is formulated as follows:

$$\begin{aligned} \max_{W_x, W_y} \quad & Trace(W_x^T C_{xy} W_y) \\ \text{subject to} \quad & W_x^T C_{xx} W_x = I; W_y^T C_{yy} W_y = I, \end{aligned} \tag{3.1}$$

where $C_{xy} = XY^T$ is the cross-correlation matrix between the two sets, and C_{xx} and C_{yy} are the auto-correlation matrices of X and Y, respectively. $W_x^T \in \Re^{p \times l}$ and $W_y^T \in \Re^{q \times l}$ are the projection matrices for X and Y, respectively and l is the number of projected dimensions. As proven in [13], W_x in Eq. (3.1) is formulated as a generalized eigenvalue decomposition problem as follows:

$$C_{xy} C_{yy}^{-1} C_{xy}^T W_x = \lambda^2 C_{xx} W_x. \tag{3.2}$$

After W_x is derived, W_y is computed by solving $C_{yy}^{-1} C_{xy}^T W_x / \lambda$. It is worth noting that CCA does not utilize any label information. As a consequence, the projected data samples using the derived projection directions W_x and W_y are not well separated. In other words, the learned space is not discriminative enough.

3.2.2 Centroid Canonical Correlation Analysis (CCCA)

A straightforward way of incorporating class label information and obtaining a discriminative low dimensional representation of two sets is to build correspondences between the centroid class vectors of the two sets, each set consists of K classes. The centroids for modalities X and Y are denoted by $U_x = [u_1^x, u_2^x, ...u_K^x]$ and $U_y = [u_1^y, u_2^y, ...u_K^y]$, respectively, where u_d^x and u_d^y are the centroids of the d^{th} class in X and Y, respectively. u_d^x and u_d^y are calculated as follows:

$$u_d^x = \frac{1}{n_{dx}} \sum_{i=1}^{n_{dx}} x_i^d, u_d^y = \frac{1}{n_{dy}} \sum_{i=1}^{n_{dy}} y_i^d, \tag{3.3}$$

where x_i^d and y_i^d are the i^{th} sample in the d^{th} class in X and Y, respectively. The variables n_{dx} and n_{dy} represent the number of the samples in the d^{th} class in X and Y, respectively. Similar to CCA, CCCA is formulated as follows:

$$\begin{aligned} \max_{W_x, W_y} \quad & Trace(W_x^T \Sigma_{xy} W_y) \\ \text{subject to} \quad & W_x^T \Sigma_{xx} W_x = I; W_y^T \Sigma_{yy} W_y = I, \end{aligned} \tag{3.4}$$

where $\Sigma_{xy} = U_x U_y^T$ is the cross-correlation matrix between the U_x and U_y, and Σ_{xx} and Σ_{yy} are the auto-correlation matrices of U_x and U_y, respectively. Similar to CCA, the above equation is solved by generalized eigenvalue decomposition. However, CCCA loses the rich information of the data because it incorporates the centroid class samples.

3.2.3 Hybrid Centroid Canonical Correlation Analysis (HCCCA)

Instead of establishing correspondences among samples in CCA or the centroid of classes in CCCA, HCCCA establishes correspondences among samples and the centroids of classes simultaneously. HCCCA aims to learn the projection basis functions, a technique which maximizes the correlation between the projected samples while maintaining a high correlation between centroid class vectors of the two sets. The HCCCA is formulated as follows:

$$\begin{aligned} \max_{W_x, W_y, \alpha_1, \alpha_2} \quad & Trace(\alpha_1 W_x^T C_{xy} W_y + \alpha_2 W_x^T \Sigma_{xy} W_y) \\ \text{subject to} \quad & W_x^T C_{xx} W_x = I; W_y^T C_{yy} W_y = I; \\ & \sum_{i=1}^{2} \alpha_i = 1; \alpha_i \geq 0, \end{aligned} \tag{3.5}$$

where C_{xy} is the cross-correlation matrix between X and Y respectively. The variable Σ_{xy} is the cross-correlation matrix between the two centroid sets U_x and U_y, respectively. The variables C_{xx} and C_{yy} are the auto-correlation matrices for X and Y, respectively. W_x and W_y are the projection matrices of X and Y, respectively. The tuning parameters $\alpha_i (i = 1, 2)$ are non-negative weights to combine the cross-correlation between X and Y and the cross-correlation between the two centroid sets U_x and U_y. The first term of the objective function $W_x^T C_{xy} W_y$ in the optimization problem maximizes the correlation between the projected samples and the second term $W_x^T \Sigma_{xy} W_y$ maximizes the correlation between the centroids of the classes in the two sets. To calculate $\alpha_i (i = 1, 2)$ in an iterative manner and with reasonable numbers, a mathematical trick from [84] [85] is adopted. A relaxation scheme is introduced by changing the tuning parameters to α_1^r and α_2^r where r is an integer and $r \geq 2$. Therefore, the HCCCA optimization problem is reformulated as follows:

$$\begin{aligned} \max_{W_x, W_y, \alpha_1, \alpha_2} \quad & Trace(W_x^T(\alpha_1^r C_{xy} + \alpha_2^r \Sigma_{xy})W_y) \\ \text{subject to} \quad & W_x^T C_{xx} W_x = I; W_y^T C_{yy} W_y = I; \\ & \sum_{i=1}^{2} \alpha_i = 1; \alpha_i \geq 0. \end{aligned} \tag{3.6}$$

Eq. (3.6) can be solved iteratively by employing an alternating optimization technique to acquire a sub-optimal solution [86]. Initially, α_1 and α_2 are fixed to compute W_x and W_y. By applying Lagrange multipliers, the following relationship are obtained as follows:

$$\begin{pmatrix} 0 & \alpha_1^r C_{xy} + \alpha_2^r \Sigma_{xy} \\ \alpha_1^r C_{xy}^T + \alpha_2^r \Sigma_{xy}^T & 0 \end{pmatrix} \begin{pmatrix} W_x \\ W_y \end{pmatrix} = \lambda \begin{pmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{pmatrix} \begin{pmatrix} W_x \\ W_y \end{pmatrix}, \tag{3.7}$$

where λ is the eigenvalues. Eq. (3.7) can be solved using generalized eigenvalue decomposition. Then, to obtain α_1 and α_2, we fix the derived W_x and W_y and apply the Lagrange multipliers method on Eq. (3.6) to update α_1 and α_2. Applying a Lagrange multipliers function on Eq. (3.6) with fixed W_x and W_y is formulated as follows:

$$\begin{aligned} L(\alpha_1, \alpha_2, \beta) = {} & \alpha_1^r Tr(W_x^T C_{xy} W_y) \\ & + \alpha_2^r Tr(W_x^T \Sigma_{xy} W_y) - \beta(\sum_{i=1}^{2} \alpha_i - 1). \end{aligned} \tag{3.8}$$

From Eq. (3.8), the derivatives with respect to α_i and β are obtained, and then the resultants are set to zero as follows:

$$\begin{aligned}
&\frac{\partial L(W_x, W_y, \alpha_1, \alpha_2, \lambda_1, \lambda_2, \beta)}{\partial \alpha_1} = r\alpha_1^{r-1} Tr(W_x^T C_{xy} W_y) \\
&- \beta = 0, \\
&\frac{\partial L(W_x, W_y, \alpha_1, \alpha_2, \lambda_1, \lambda_2, \beta)}{\partial \alpha_2} = r\alpha_2^{r-1} Tr(W_x^T \Sigma_{xy} W_y) \\
&- \beta = 0, \\
&\frac{\partial L(W_x, W_y, \alpha_1, \alpha_2, \lambda_1, \lambda_2, \beta)}{\partial \beta} = \sum_{i=1}^{2} \alpha_i - 1 = 0.
\end{aligned} \tag{3.9}$$

The parameters α_1 and α_2 are calculated and updated by Eq. (3.10) and Eq. (3.11):

$$\alpha_1 = \frac{(\frac{1}{Tr(W_x^T C_{xy} W_y)})^{\frac{1}{r-1}}}{(\frac{1}{Tr(W_x^T C_{xy} W_y)})^{\frac{1}{r-1}} + (\frac{1}{Tr(W_x^T \Sigma_{xy} W_y)})^{\frac{1}{r-1}}}. \tag{3.10}$$

$$\alpha_2 = \frac{(\frac{1}{Tr(W_x^T \Sigma_{xy} W_y)})^{\frac{1}{r-1}}}{(\frac{1}{Tr(W_x^T C_{xy} W_y)})^{\frac{1}{r-1}} + (\frac{1}{Tr(W_x^T \Sigma_{xy} W_y)})^{\frac{1}{r-1}}}. \tag{3.11}$$

After that, α_1 and α_2 are fixed to compute the new W_x and W_y. The projection matrices W_x and W_y and the tuning parameters α_1 and α_2 are updated in an iterative manner until they are converged. The algorithm for solving the HCCCA optimization problem is summarized in Algorithm 3.1.

However, HCCCA can fuse two modalities, this leads us to generalize HCCCA to a multimodal fusion handling more than two modalities.

Algorithm 3.1 HCCCA Algorithm

```
Require: Two sets $X$ and $Y$, the class centroids of the two sets $U_x$ and $U_y$, and $r \geq 2$
Ensure: $W_x$, $W_y$, $\alpha_1$, and $\alpha_2$
1: Compute $C_{xy}$, $\Sigma_{xy}$, $C_{xx}$, and $C_{yy}$
2: Initialize $\alpha_1 = \alpha_2 = \frac{1}{2}$
3: Repeat
4:   Obtain $W_x$ and $W_y$ using Eq. (3.7)
5:   Compute $\alpha_1$ and $\alpha_2$ using Eq. (3.10) and Eq. (3.11)
6: Until Convergence
```

3.2.4 Multimodal Hybrid Centroid Canonical Correlation Analysis (MHCCCA)

In order to obtain the shared space for two or more data modalities or sets, we extend HCCCA to MHCCCA by following steps similar to MCCA [14]. Given P sets of data samples $X_1, ...X_P$, with dimensions $m_1, ...m_P$, the objective of MHCCCA is to maximize the correlation of data samples from different heterogeneous modalities, while maintaining a high correlation of the centroids of the classes of the different modalities. The optimization problem is formulated as follows:

$$\begin{aligned} \max_{W_1,..W_P,\alpha_1,\alpha_2} \quad & Trace(\sum_{k,j=1,k\neq j}^{P} W_k^T A_{kj} W_j) \\ \text{subject to} \quad & \sum_{i=1}^{P} W_i^T C_{ii} W_i = P(I); \sum_{l=1}^{2} \alpha_l = 1; \alpha_l \geq 0. \end{aligned} \tag{3.12}$$

The matrix A_{kj} is of dimension $m_k \times m_j$ and is obtained as $A_{kj} = \alpha_1^r C_{kj} + \alpha_2^r \Sigma_{kj}$, $C_{kj} = X_k Y_j^T$ and $\Sigma_{kj} = U_k U_j^T$. The matrix C_{ii} is the auto-correlation matrix of the i^{th}

modality or set. The above optimization problem can be rewritten as follows:

$$\max_{W_1 \dots W_P, \alpha_1, \alpha_2} \quad Trace\left(\begin{pmatrix} W_1 & W_2 & \cdots & W_P \end{pmatrix} \begin{pmatrix} 0_{1,1} & A_{1,2} & \cdots & A_{1,P} \\ A_{2,1} & 0_{2,2} & \cdots & A_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ A_{P,1} & A_{P,2} & \cdots & 0_{P,P} \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_P \end{pmatrix}\right) \tag{3.13}$$

$$\text{subject to} \quad \begin{pmatrix} W_1 & W_2 & \cdots & W_P \end{pmatrix} \begin{pmatrix} C_{1,1} & 0_{1,2} & \cdots & 0_{1,P} \\ 0_{2,1} & C_{2,2} & \cdots & 0_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{P,1} & 0_{P,2} & \cdots & C_{P,P} \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_P \end{pmatrix} = P(I); \sum_{i=1}^{2} \alpha_i = 1; \alpha_i \geq 0$$

$$A = \begin{pmatrix} 0 & A_{1,2} & \cdots & A_{1,P} \\ A_{2,1} & 0 & \cdots & A_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ A_{P,1} & A_{P,2} & \cdots & 0 \end{pmatrix}; C = \begin{pmatrix} C_{1,1} & 0 & \cdots & 0 \\ 0 & C_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C_{P,P} \end{pmatrix}, \tag{3.14}$$

where the matrices A and C are of dimension $N \times N$ and $N = \sum_{i=1}^{P} m_i$. First, α_1 and α_2 are initialized and fixed to compute W. By applying Lagrange multipliers method, we obtain the following relationship

$$PAW = \lambda CW, \tag{3.15}$$

where $W = [W_1^T, W_2^T, ...W_P^T]^T$ is the projection matrix which is associated with the P modalities or sets. Then, using the derived projection matrix W and the tuning parameters (α_1, α_2) are updated by Eq. (3.16) and Eq. (3.17):

$$\alpha_1 = \frac{(\frac{1}{Tr(V)})^{\frac{1}{r-1}}}{(\frac{1}{Tr(V)})^{\frac{1}{r-1}} + (\frac{1}{Tr(Q)})^{\frac{1}{r-1}}}, \tag{3.16}$$

$$\alpha_2 = \frac{(\frac{1}{Tr(Q)})^{\frac{1}{r-1}}}{(\frac{1}{Tr(V)})^{\frac{1}{r-1}} + (\frac{1}{Tr(Q)})^{\frac{1}{r-1}}}, \tag{3.17}$$

where $V = \sum_{k,j=1,k\neq j}^{P} W_k^T C_{kj} W_j$ and $Q = \sum_{k,j=1,k\neq j}^{P} W_k^T \Sigma_{kj} W_j$. After that, α_1 and α_2 are fixed to compute the new W , using a procedure similar to a HCCCA until they converge. The algorithm for solving MHCCCA optimization problem is shown in Algorithm 3.2.

Algorithm 3.2 MHCCCA Algorithm

```
Require: $P$ sets $X_{i=1:P}$,the class centroids of the $P$ sets $U_{i=1:P}$, and $r \geq 2$
Ensure: $W$, $\alpha_1$, and $\alpha_2$
1: Compute $A$ and $C$ using Eq. (3.14)
2: Initialize $\alpha_1 = \alpha_2 = \frac{1}{2}$
3: Repeat
4:   Obtain $W$ using Eq. (3.15)
5:   Compute $\alpha_1$ and $\alpha_2$ using Eq. (3.16) and Eq. (3.17)
6: Until Convergence
```

3.3 Descriptors

RGB videos, depth videos, skeleton data, and accelerometer data are rich sources of human movements and activities. One of the main challenges in human action recognition is feature extraction. In this section, a new descriptor for skeleton representation using Bag of Angles (BoA) is investigated. Additionally, Hierarchical Pyramid DMM (HP-DMM) for depth videos, MHI and MEI for RGB videos, and Statistical Features (SF) for accelerometer data are employed.

3.3.1 Skeleton Data

In early work, Johansson *et al.*, claimed that skeleton data is discriminative enough to identify human gestures [87]. Here, a new descriptor for skeleton data named BoA is proposed [88]. For each frame, 70 features are computed following the feature extraction from [89]. The first 35 features are angles defined by triplets of joints describing the posture at a time instant. For example, the angle between the right elbow, the neck, and the left elbow describes the posture of the two arms with respect to the neck at a time instant. Figure 3.2 shows examples for the angle between joints. On the other hand, the second set of 35 features are the joint angle accelerations describing the dynamic trajectory of the angle joints over time. So, the proposed 70 features describe both the posture and dynamics of skeleton data. The calculated angles are considered as local features. To properly represent the locally extracted features, BoA model is computed. First, the codebook for the features is generated. K-means is adopted to cluster the space into regions, where each region represents a codeword. Then, the local descriptors are encoded using vector quantization (VQ) with 50 codewords. In VQ, each descriptor votes for only one codeword. Finally, a histogram is computed as a representation for each video. The pipeline of BoA is illustrated in Figure 3.3.

3.3.2 Depth Video

Depth videos are important in action recognition because they capture 3D action structure effectively. Here, we adopt Hierarchical Pyramid Depth Motion Map (HP-DMM) [90] as the feature for depth videos which is an improved representation over Depth

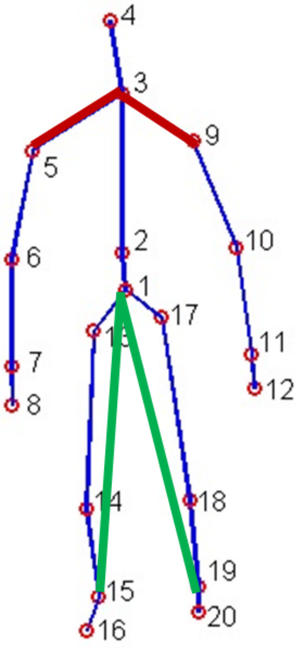


Figure 3.2: Illustration of angles formed by three joints. The red angle between the right shoulder, the neck, and the left shoulder describes the shoulder's posture with respect to the neck at a time instant. The green angle between the right ankle, the center hip, and the left ankle describes the legs' posture with respect to the center hip at a time instant.

Motion Map (DMM) for addressing the challenge of self-occlusion. To construct HP-DMM, the video sequence is divided into multiple video sections as illustrated in Figure 3.4. For each video section, DMM is computed as follows:

$$DMM_{F,S,T} = \sum_{i=1}^{N-1} |DM^{i+1}_{F,S,T} - DM^{i}_{F,S,T}|, \tag{3.18}$$

where i represents the frame index number, N is the total number of frames in a video sequence, and F, S, and T are the front, side and top projections of the depth maps, respectively. HP-DMM has the ability to capture more detailed information about the action and the fine changes of human movements since it captures the sub-actions within

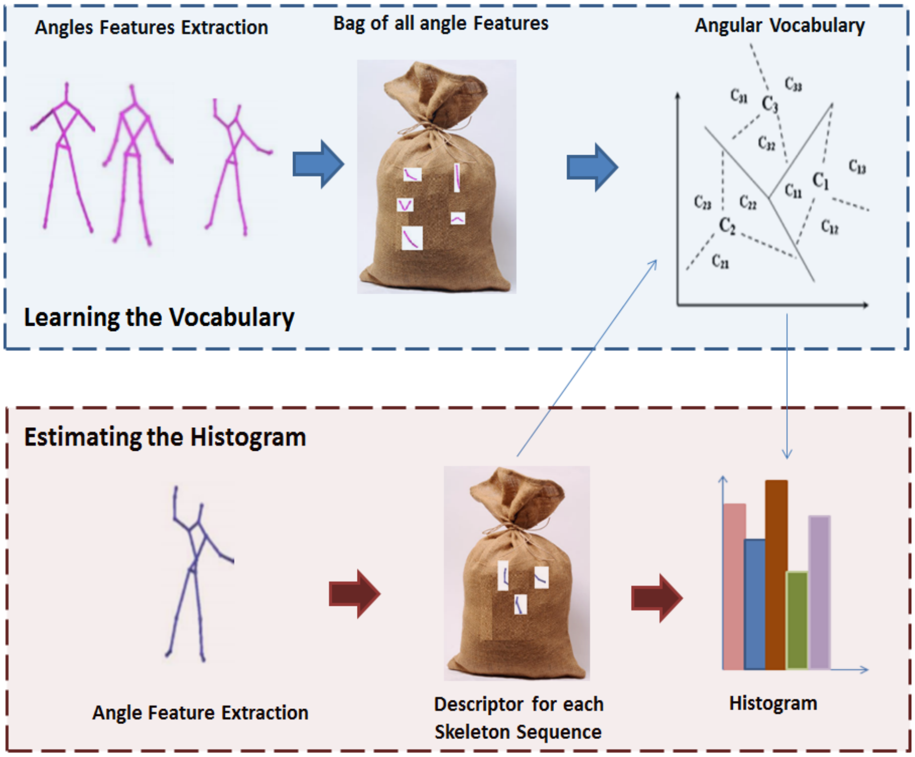


Figure 3.3: The pipeline of BoA descriptor. The upper part represents vocabulary learning and the lower part represents histogram estimation.

the video sequence. Two-level pyramids with two partitions, which have previously shown good results [90] are employed. The construction of a hierarchical pyramid with two levels is shown in Figure 3.5. Finally, HOG [31] is used to describe the local shape of each $DMM_{F,S,T}$ in each partition. The final descriptor is the concatenation of HOG descriptors of each DMM.

3.3.3 RGB Video

Motion History Image MHI and Motion Energy Image MEI are adopted as RGB videos descriptors [4]. MHI is used to represent the recency of motion in a video sequence. On the other hand, MEI is a binary image used to represent the presence of motion. Instead of using Sobel gradients mask as in [91], HOG is used to capture the local

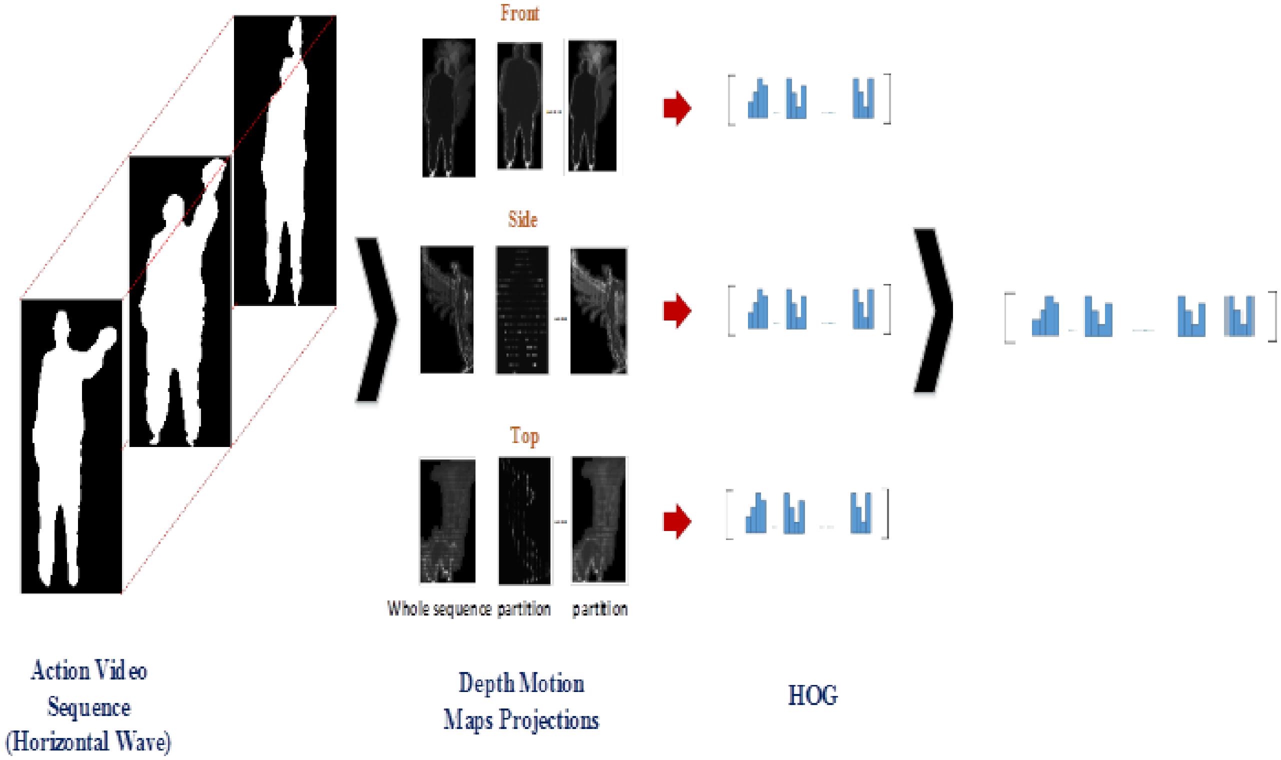

Front
Side
Top
Whole sequence partition
partition
Action Video Sequence (Horizontal Wave)
Depth Motion Maps Projections
HOG

Figure 3.4: The generation of HP-DMM descriptor.

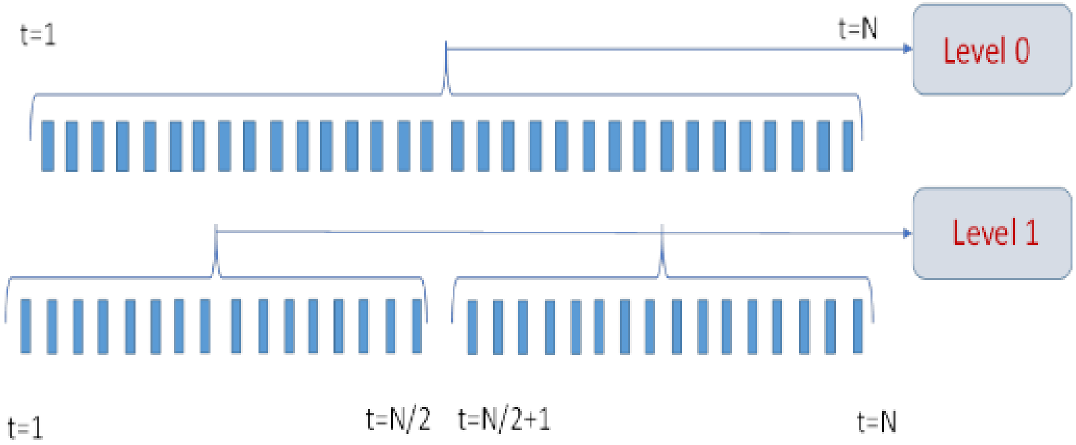


Figure 3.5: Hierarchical pyramid of two levels with two partitions.

Table 3.1: Statistical Features.

Feature	Definition %
Mean	$mean(x) = \frac{1}{N} \sum_{n=1}^{N} x[n]$
Max	The highest value in the set of data samples.
Min	The smallest value in the set of data samples.
Variance	$\sigma_x^2 = \overline{x^2} - (\overline{x})^2$

shape of each image. The final descriptor is the concatenation of the HOG descriptors for both images.

3.3.4 Accelerometer Data

The fourth modality is the data from an accelerometer. The accelerometer provides the acceleration measurements along the three orthogonal Cartesian axes. Many features can be extracted from the accelerometer data to represent actions such as Statistical Features (SF), energy features and frequency domain features [92]. SF are used as features for accelerometer data due to their simplicity, effectiveness, and good performance [93]. Such features can discriminate high energy actions from low energy actions. Table 3.1 shows the list of SFs with brief definitions. Each axis of the three orthogonal axes is partitioned into N_S temporal windows. In each temporal window, mean, max, min, and variance are computed for each temporal window. The main motivation of computing SFs for each temporal window is to obtain quantitative measurements that allow signal discrimination. The number of temporal windows N_S affects the quality of the features. An improper number of temporal windows N_S leads to non-discriminative features. The final descriptor is the concatenation of the four SFs of each temporal window for each axis of the three orthogonal axes.

3.4 Experimental Results

In order to evaluate the effectiveness of the proposed framework, comprehensive experiments are conducted on four publicly available multimodal human action datasets. In this section, datasets description, data visualization, and performance evaluations are presented.

3.4.1 Multimodal Action Datasets

In the evaluation, the experiments are conducted on four different action datasets: MSR Action3D [94], UTD Multimodal Human Action Dataset (UTD-MHAD) [95], UTD-MHAD-Kinect V2 [96], and Berkeley MHAD [80] which incorporate different modalities for the the task of human action recognition.

The **MSR Action3D Dataset** contains 1114 action sequences (557 depth sequence, and 557 skeleton sequences) for 20 actions performed by 10 subjects. All the sequences were captured with Kinect sensor. The 20 actions are: *high wave, horizontal wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pickup throw.* Each subject performed every action two or three times. We followed the same settings in [78], where all the 20 actions were employed. Half of the subjects were used for training and the rest for testing.

The **UTD Multimodal Human Action Dataset** (UTD-MHAD) is composed of 3444 sequences (861 RGB video sequences, 861 depth video sequences, 861 skeleton sequences, and 861 accelerometer data samples) for 27 actions performed by 8 subjects. RGB sequences, depth sequences, and skeleton sequences were captured using Kinect sensor. Accelerometer data sequences were captured by a wearable inertial sensor, which was worn on subject's right wrist or right thigh depending on whether the action was mostly an arm or a leg type action. The 27 actions are: *right arm swipe to the left, right* *arm swipe to the right, right hand wave, two hand front clap, right arm throw, cross arms in the chest, basketball shoot, right hand draw x, right hand draw circle, right hand draw circle, draw triangle, bowling, front boxing, baseball swing from right, tennis right hand forehand swing, arm curl, tennis serve, two hand push, right hand knock on door, right hand catch an object, right hand pick up and throw, jogging in place, walking in place, sit to stand, stand to sit, forward lunge, and squat.* Each subject performed every action four times. We followed the experimental settings as in [95], where all the 27 actions were employed. Half of subjects were used for training and the rest for testing.

The **UTD-MHAD-Kinect V2 Dataset** consists of 1200 sequences (400 depth video sequences, 400 skeleton sequences, and 400 accelerometer data samples) for 10 actions performed by 6 subjects. Depth sequences, and skeleton sequences were captured using a Kinect V2 sensor. Accelerometer data sequences were captured by a wearable inertial sensor, which was worn on subject's right wrist. The 10 actions are: *right hand high wave, right hand catch, right hand high throw, right hand draw X, right hand draw tick, right hand draw circle, right hand horizontal wave, right hand forward punch, right hand hammer, and Hand clap.* Each subject performed each action five times. Half of subjects were used for training, and the others for testing.

The **Berkeley MHAD Dataset** is composed of 11 actions performed by 12 subjects. Each subject performed every action five times. The dataset was captured using two stereo cameras, two quad cameras, two Kinects, six accelerometer sensors, eight motion capture sensors. The 11 actions are: *jumping in place, jumping jacks, bending - hands up all the way down, punching, waving - two hands, Waving-one hand, clapping hands, throwing a ball, sit down then stand up, sit down, and stand up.* In the experiments, three modalities are used: RGB camera, depth video sequences, and accelerometer data. Half of subjects are dedicated for training and the rest for testing. For RGB video sequences and depth sequences, the video sequences were down sampled to 120×160 after foreground extraction.

3.4.2 Analytical Evaluation of BoA

In order to verify the performance achieved by the proposed descriptor BoA, we conducted experiments on the MSR Action3D dataset by comparing the performance against different baseline skeleton descriptors. As shown in Table 3.2, BoA displays the best performance in comparison. The evaluation results indicate that the proposed BoA descriptor is more effective due to its ability to capture posture and dynamic trajectory of the joints over time.

Table 3.2: Recognition Accuracy comparison between BoA against other baseline skeleton descriptors on MSR Action3D Dataset.

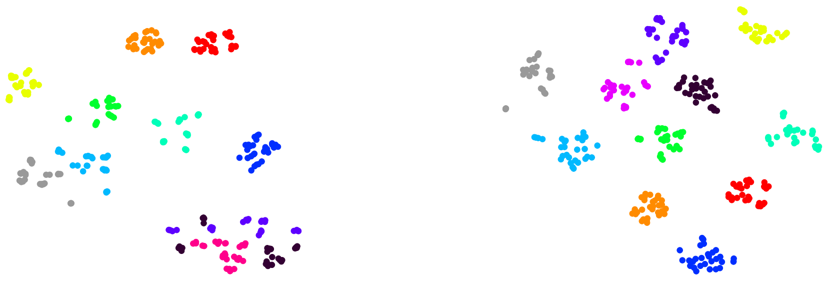
Skeleton Descriptor	Recognition Accuracy %
HOJ3D [68]	79.0
EigenJoints [10]	83.3
HOD [11]	85.5
COV3D [71]	84
BoA	**86.9**

3.4.3 Visualization of MHCCCA

to gain qualitative insight into the feature representations learned from MHCCCA, a visualization experiment was conducted. The visualization of MCCA and MHCCCA common feature spaces were constructed. For visualization, we used the t-distributed Stochastic Neighbor Embedding (t-SNE) [97] to reduce the high-dimensional space into 2D space. Berkeley MHAD dataset was chosen for this task due to the fact that this dataset contains actions which are quite similar. The results are reported in Figure 3.6, where it is observed that the between-class distance is undesirably small in MCCA. On the other hand, the proposed MHCCCA has a much larger distance and separates all the classes properly. As evidenced by the experiment, it can be concluded that the fusion using MHCCCA has a better capability of learning a discriminative shared space than CCA.

3.4.4 Performance Evaluation

The experimental results are reported on the four datasets: MSR Action3D dataset, UTD Multimodal Human Action Dataset, UTD-MHAD-Kinect V2 dataset, and Berkeley MHAD dataset. The effectiveness of MHCCCA was verified by comparing its performance against methods based on similar principles in terms of recognition accuracy. The first 100 dimensions (90 % of total covariance) of the projected features in the common space were selected, and then the maximum recognition accuracy from these 100 dimensions were picked. The projected features in the shared common space were concatenated and Linear SVM was used for classification [98].



(a) t-SNE embedding of MCCA. (b) t-SNE embedding of MHCCCA.

Figure 3.6: The visualization of feature space comparison between (a) MCCA and (b) MHCCCA for individual feature.

MSR Action3D dataset

First, the recognition accuracy for each modality is studied. As the MSR Action3D dataset has only depth and skeleton,HP-DMM and BoA recognition performances are evaluated and the results are given in Table 3.3; which shows that HP-DMM and BoA have a recognition accuracy of 91.92 % and 86.92 %, respectively. Table 3.4 illustrates the recognition accuracy of MHCCCA, CCCA, and CCA by fusing HP-DMM and BoA features. It is observed that MHCCCA outperforms CCCA and CCA demonstrating the effectiveness of the proposed method on MSR Action3D dataset. MHCCCA has a maximum recognition accuracy of 93.5 %, which is higher than CCA and CCCA by nearly 2.3 % and 5 %, respectively. The comparison indicates that MHCCCA is able to learn a more discriminative common feature space by using the label information of the training data.

Table 3.3: Recognition Accuracy for each Modality (Depth and Skeleton) on MSR Action3D dataset.

Modality	Recognition Accuracy %
HP-DMM	91.92
BoA	86.92

The proposed method is compared with a number of existing methods. Table 3.5 demonstrates that the proposed method outperforms the existing methods under the same experimental settings.

Table 3.4: Recognition accuracy and projected dimensions comparison on MSR Action3D dataset.

Multimodal Learning	Recognition Accuracy %
CCA	91.2
CCCA	88.1
MHCCCA	93.5

Table 3.5: Comparison Between the Proposed Human Action Recognition Framework and Some of the Existing Methods on MSR Action3D dataset.

Methods	Recognition Accuracy %
DMM [78]	86.54
HON4D [65]	88.9
Actionlet ensemble [78]	88.2
Vemulapalli *et al.* [99]	89.5
Tran *et al.* [66]	91.9
DMM-LBP-FF *et al.* [64]	91.9
Proposed	**93.5**

UTD-MHAD dataset

First, the recognition accuracy for each modality was then computed. The recognition accuracies for HP-DMM, BoA, MHI, and SF were first computed. Experiment was conducted on the accelerometer data to find the optimum number of temporal windows. Figure 3.7 shows recognition accuracy for different numbers of temporal windows. Each sequence is divided into N_S temporal windows and SFs are calculated for each window. From this figure, the number of temporal windows N_S was chosen to be 15 with an accuracy of 76.51 %. Table 3.6 shows the recognition accuracy for each modality, where it can be seen that BoA has recognition accuracy of 85.35 % which is higher than the other modalities.

Table 3.6: Recognition Accuracy for each Modality (Depth, Skeleton, RGB, and Accelerometer data) on UTD-MHAD dataset.

Modality	Recognition Accuracy %
HP-DMM	73.72
BoA	85.35
MHI	73.26
SF	76.51

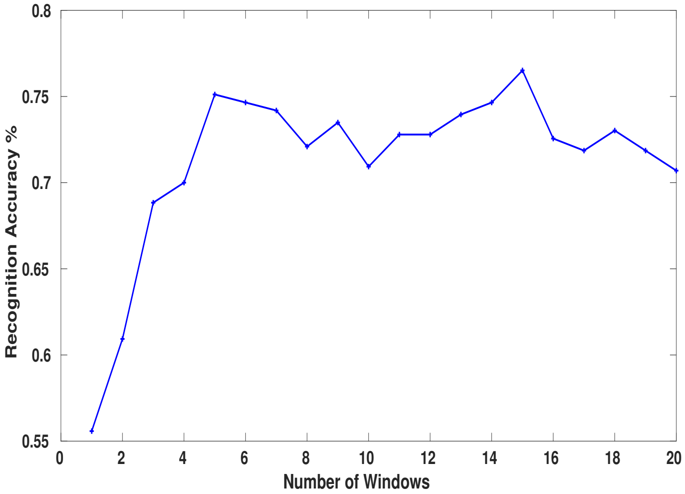


Figure 3.7: Human Action Recognition accuracy of SF with the number of temporal windows on UTD-MHAD dataset.

To exploit the benefits of multimodal learning using MHCCCA for two modalities, experiments were conducted on all possible combinations of two modalities and presented in Table 3.6.

Table 3.7: Recognition Accuracy Comparison among MHCCCA, CCCA, and CCA on UTD-MHAD dataset.

Modalities	CCA	CCCA	MHCCCA
HP-DMM, BoA	83.3	81.4	84.6
HP-DMM, SF	81.2	79.9	81.4
HP-DMM, MHI	80.0	80.5	83.5
BoA, SF	79.3	80.7	80.8
BoA, MHI	84.2	84.2	86.5
MHI, SF	79.53	76.74	79.53

The fusion of BoA and MHI has the highest recognition accuracy. The recognition accuracy of MHCCCA (BoA, MHI) has recognition accuracy of 86.5 %, which is higher than CCA and CCCA by more than 2 %. To exploit the benefits of fusion using MHCCCA for three modalities, extensive experiments were conducted on all possible combinations of three modalities and presented in Table 3.8, from which we can see that, the fusion of MHI, BoA and SF has the highest recognition accuracy. The accuracy of MHCCCA (MHI, BoA, SF) has recognition accuracy of 95.8 % which is higher than MCCA and MCCCA by 2 % and 3 %, respectively.

Table 3.8: Recognition Accuracy Comparison among MHCCCA, Multimodal Centroid Canonical Correlation Analysis (MCCCA), and MCCA on UTD-MHAD dataset.

Modalities	MCCA	MCCCA	MHCCCA
MHI, BoA, SF	93.9	92.8	95.8
MHI, HP-DMM, SF	92.1	91.2	94.4
HP-DMM, BoA, SF	94.4	93.3	94.2
HP-DMM, MHI, BoA	91.4	90.9	92.0

Additionally, this combination (MHI, BoA, SF) has the highest accuracy compared to other combinations. It can be clearly noticed that the fusion of three modalities, using MHCCCA, has higher performance than only two modalities. Then, the fusion of the four features sets: BoA, HP-DMM, MHI and SF, is exploited where the performance of MHCCCA is compared to those of MCCA and MCCCA in Table 3.9. The results demonstrate that the proposed method outperforms the other methods under the same experimental settings.

Table 3.9: Recognition accuracy and projected dimensions comparison on UTD-MHAD dataset.

Multimodal Learning	Recognition Accuracy %
MCCA	94.6
MCCCA	94.2
MHCCCA	96.1

In Table 3.10, the proposed method is compared against some of the existing methods on UTD-MHAD dataset. The results indicate the superiority of the proposed against some of the other existing methods.

Table 3.10: Comparison Between the Proposed Human Action Recognition Framework and Some of the Existing Methods on UTD-MHAD dataset.

Methods	Recognition Accuracy %
Kinect-Inertial [95]	79.1
SOS [100]	86.97
JTM [17]	85.81
Proposed	**96.1**

UTD-MHAD-Kinect V2 dataset

First, we look into the recognition accuracy for each modality in UTD-MHAD-Kinect V2 dataset. The recognition performances of the three modalities, HP-DMM, BoA, and SF were first obtained. To obtain SF, experiments were conducted on the accelerometer data to find the optimum number of temporal windows. Each sequence is divided into N_S temporal windows and SF are calculated for each temporal window. Figure 3.8 shows the recognition accuracy with the number of temporal windows. From the figure, the number of temporal windows $N_S = 8$ is chosen. Table 3.11 illustrates the recognition accuracy for each modality, showing that SF has the highest recognition accuracy, 86 %, among the features compared.

Table 3.11: Recognition Accuracy for each Modality (Depth, Skeleton, and Accelerometer data) on UTD-MHAD-Kinect V2 dataset.

Modality	Recognition Accuracy %
HP-DMM	62
BoA	84
SF	86

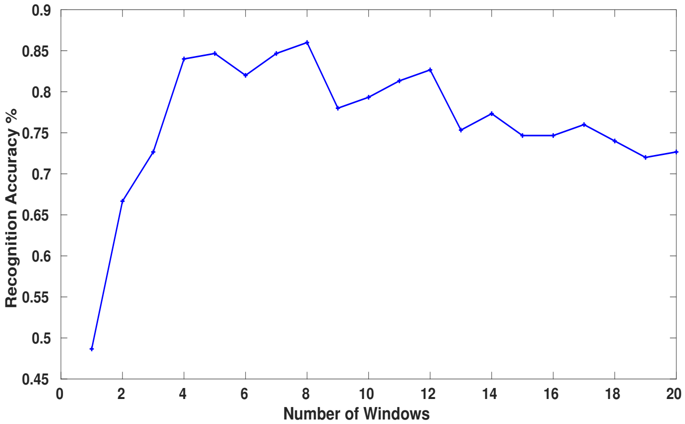


Figure 3.8: Human Action Recognition accuracy of Statistical Features (SF) with the number of temporal windows on UTD-MHAD-Kinect V2 dataset.

To evaluate MHCCCA, a comparison is conducted with all possible of two modalities fusion combinations with the recognition accuracy shown in Table 3.12.

Table 3.12: Recognition Accuracy Comparison among the Proposed Fusions Techniques MHCCCA, CCCA, and CCA on UTD-MHAD-Kinect V2 dataset.

Modalities	CCA	CCCA	MHCCCA
HP-DMM, BoA	82	80	83.3
HP-DMM, SF	86.7	60	86
BoA, SF	84	82.7	86.7

From Table 3.12, the recognition accuracy of MHCCCA (BoA, SF) has the highest recognition accuracy of 86.7 %. Also, CCA (HP-DMM, SF) has 86.7 % as well. Then, comparison among MCCA, MCCCA and MHCCCA with the three features as the input is conducted. The results show that the proposed method MHCCCA achieves 90 % which is higher than MCCCA and MCCA by 2 % and 7.5 %. Table 3.13 summarizes the results on UTD-MHAD-Kinect V2 dataset.

Table 3.13: Recognition accuracy and projected dimensions comparison on UTD-MHAD-Kinect V2 dataset.

Multimodal Learning	Recognition Accuracy %
MCCA	82.7
MCCCA	88
MHCCCA	90

Berkeley MHAD dataset

Again, the recognition accuracy of each modality was first studied. Experiments were conducted on the accelerometer data to find the optimum number of temporal windows of the SF. Each sequence is divided into N_S temporal windows and SF are calculated for each temporal window. From Figure 3.9, the number of temporal windows is chosen such that $N_S = 13$.

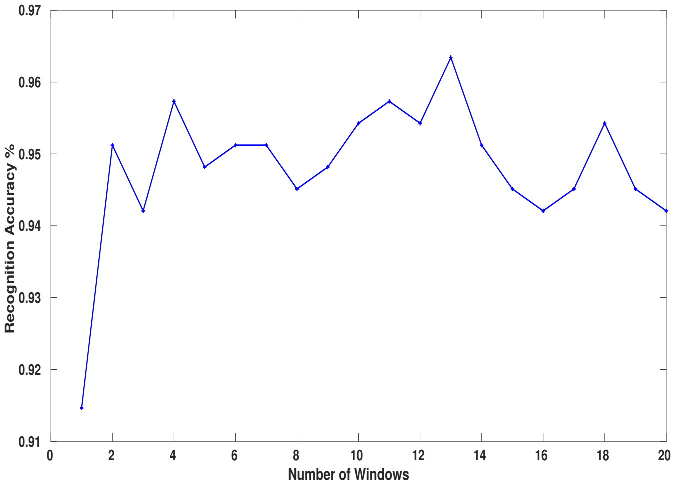


Figure 3.9: Human Action Recognition accuracy of Statistical Features (SF) with the number of temporal windows on Berkeley MHAD dataset.

Table 3.14 reports the recognition accuracy for each modality. It shows, MHI has the highest recognition accuracy, 96.65 % among the three modalities.

Table 3.14: Recognition Accuracy for each Modality (Depth, RGB, and Accelerometer data) on Berkeley MHAD dataset.

Modality	Recognition Accuracy %
HP-DMM	93
MHI	96.65
SF	96.34

Table 3.15 compares the recognition accuracy of MHCCCA, CCCA, and CCA by

fusing the all possible combinations.

Table 3.15: Recognition Accuracy Comparison among MHCCCA, CCCA, and CCA on Berkeley MHAD dataset.

Modalities	CCA	CCCA	MHCCCA
HP-DMM, MHI	96	86.6	96.4
HP-DMM, SF	98	92.7	98.2
MHI, SF	98.2	91.2	98.8

From the table, it is observed that the combination (MHI, SF) has the highest recognition accuracy of 99.09 % higher than CCA and CCCA by nearly 0.6 % and 7 %, respectively.

The performance of MHCCCA (HP-DMM, MHI, SF) against MCCCA (HP-DMM, MHI, SF) and MCCA (HP-DMM, MHI, SF) are then evaluated and the results summarized in Table 3.16. The results show that MHCCCA (HP-DMM, MHI, SF) outperforms all other techniques with recognition accuracy of 99.8 % which is higher than MCCA (HP-DMM, MHI, SF) and MCCCA (HP-DMM, MHI, SF) by 1.3 % and 7.4 %, respectively.

Table 3.16: Recognition accuracy and projected dimensions comparison on Berkeley MHAD dataset.

Multimodal Learning	Recognition Accuracy %
MCCA	98.5
MCCCA	92.4
MHCCCA	99.8

3.5 Conclusion

In this chapter, a human action recognition framework with different modalities based on Multimodal Hybrid Centroid Canonical Correlation Analysis (MHCCCA) for fusing two or more modalities was presented. The proposed MHCCCA find the discriminative common space with the knowledge of the class labels from the training dataset. Additionally, a novel skeleton feature representation using Bag of Angles (BoA) proposed. The experimental results show the effectiveness of the proposed framework and the ability of and MHCCCA in discovering the discriminative common space and classifying the similar actions with same testing time complexity of CCA.

In addition, the learned common space using MHCCCA and MCCA are visualized where MHCCCA can learn a more discriminative common space than MCCA.

Chapter 4

Action Recognition via MGLPCCA

4.1 Introduction

In this chapter, the problem of human action recognition is studied, in which each action is captured by multiple sensors and represented by multisets of features. In the previous chapter, the performance of the handcrafted features is constrained by their limited capacity, leading to unsatisfactory results on MSR-Action 3D datasets. Authors in [101] studied the fusion of handcrafted and deep learning features for emotion recognition and their results demonstrated the complementarity between handcrafted and deep learning features. Such an observation motivated us to migrate from handcrafted features to a mixture of handcrafted and deep learning features in order to boost the performance of action recognition on small to medium scale datasets using information fusion technique. Herein, deep learning based features are incorporated in the framework to improve the quality of feature representations. The deep learning based features take advantage of Deep ConvNets trained on large scale datasets, as a feature representation for small scale datasets [102]. First, we introduce Globality Locality Preserving Canonical Correlation Analysis (GLPCCA) to learn the common feature subspace and preserve the locality and globality of the features in the shared space, however it is only able to fuse two sets. Therefore, it is generalized to Multiset Globality Locality Preserving Canonical Correlation Analysis (MGLPCCA), which aims to deal with two or more sets. The proposed MGLPCCA is able to learn a low-dimensional common subspace that preserves the local and global structures of data samples. The reasons for introducing the proposed MGLPCCA are:

- MGLPCCA is capable of discovering the manifold structure that preserves both the local and global structures of the original feature space.
- The locality preserving improves the discriminative power of the features in the learned subspace [103].
- The globality preserving improves discriminative power of the projected features in the learned subspace [104].
- Based on Globality Preserving Projections (GPP), MGLPCCA further considers the relationship among various sets.

4.2 Review on GPP

The core concept of Locality Preserving Projections (LPP) algorithms [105] is to encode the manifold structure of the samples using a Laplacian graph constructed in either a supervised or unsupervised manner. In the case of supervised, the Laplacian graph captures only local structure and ignores the global structure which is important in distinguishing between classes. To overcome this problem, Globality Locality Preserving Projections (GLPP) was introduced in [106]. GLPP captures both the local and the global structures of the data samples as it considers the manifold as an interaction between the local (intra-class) and global (inter-class) structures. Let $X \in \Re^{p \times n}$ be the set of data samples. GLPP aims to learn the projection matrix W that transforms the data samples into a subspace, where the geometric relations of the projected data samples are well preserved. Let $U = [u_1, ...u_d, ...u_K]$ denotes the mean space of data samples, where u_d is the mean of the data samples of the class d and K is the number of classes. The variable U is used for preserving the global structure of the whole data. A reasonable criterion for revealing the real relationship among data samples is to learn local and global Laplacian graphs, which are formulated as follows:

$$\Phi_{inter} = \frac{1}{2}\sum_{i,j}(W^T u_i - W^T u_j)^2 B_{ij}, \tag{4.1}$$

$$\Phi_{intra} = \frac{1}{2}\sum_{c\in C_{set}}\sum_{i,j\in c}(W^T x_i - W^T x_j)^2 S_{ij}, \tag{4.2}$$

where C_{set} is the set of classes, c is the set of data samples which belongs to the same class, and S_{ij} and B_{ij} are the the adjacency weights of intra-class and inter-class data samples. The ij^{th} element S_{ij} and B_{ij} in adjacency weight matrices S and B are defined as follows:

$$S_{ij} = \begin{cases} exp(-\|x_i - x_j\|^2/t_S) & i,j \in c, c \in C_{set}, i \neq j \\ 0 & otherwise \end{cases}, \tag{4.3}$$

$$B_{ij} = \begin{cases} exp(-\|u_i - u_j\|^2/t_B) & i \neq j \\ 0 & otherwise \end{cases}. \tag{4.4}$$

The parameter t_S is chosen as the mean square distance $\sum_{i,j}(\|x_i - x_j\|^2/(n(n-1)))$ as in [107]. Similarly, the parameter t_B is chosen as the mean square distance $\sum_{i,j}(\|u_i - u_j\|^2/(K(K-1)))$. The objective function for GLPP is to minimize the following objective function:

$$\psi = \Phi_{inter} + \beta\Phi_{intra} = \frac{1}{2}\sum_{i,j}(W^T u_i - W^T u_j)^2 B_{ij} + \frac{1}{2}\beta \sum_{c \in C_{set}} \sum_{i,j \in c}(W^T x_i - W^T x_j)^2 S_{ij} \tag{4.5}$$

where β is a tuning parameter to control the contribution of preserving local and global structure. Eq. (4.5) can be rewritten as follows:

$$\psi = W^T U(G - B)U^T W + \beta \sum_{c \in C} W^T X_c(D - S)X_c^T W, \tag{4.6}$$

where G is a diagonal matrix and its entries are row sums of B, $G_{ii} = \sum_j B_{ij}$. D is a diagonal matrix and its entries are row sum of S, $D_{ii} = \sum_j S_{ij}$. Both S_{ij} and B_{ij} are calculated using Eq. (4.3) and Eq. (4.4), respectively. Therefore Eq. (4.6) can be reduced to:

$$\psi = W^T(UHU^T + \beta \sum_{c \in C} X_c L X_c^T)W, \tag{4.7}$$

where $H = G - B$ and $L = D - S$ are the Laplacian graph of the intra-class and inter-class, respectively. The minimization of Eq. (4.7) with respect to W is given in the following form:

$$\min_{W} \quad Trace(W^T A W), \tag{4.8}$$

where $A = UHU^T + \beta XLX^T$ is a positive semi definite matrix. Therefore, the above problem can be solved using eigenvalue decomposition.

4.3 The Proposed Globality Locality Preserving Canonical Correlation Analysis (GLPCCA)

GLPCCA aims to preserve both the local and global structures while maximizing the correlation between the two data features or sets. In other words, GLPCCA is CCA which incorporates the local and global information in order to enrich the feature representation, and can be formulated as follows:

$$\begin{aligned} \max_{W_x, W_y} \quad & Trace(W_x^T (U_x H_{xy} U_y^T + \beta X L_{xy} Y^T) W_y) \\ \text{subject to} \quad & W_x^T (U_x H_{xx} U_x^T + \beta X L_{xx} X^T) W_x = I; \\ & W_y^T (U_y H_{yy} U_y^T + \beta Y L_{yy} Y^T) W_y = I, \end{aligned} \tag{4.9}$$

where W_x and W_y are the projection matrices for X and Y, and $U_x = [u_1^x, ..u_i^x, ..u_K^x]$ and $U_y = [u_1^y, ..u_i^y, ..u_K^y]$ are the mean class sample spaces for X and Y, respectively.

$L_{xx} = D_{xx} - S_x \circ S_x$, $L_{yy} = D_{yy} - S_y \circ S_y$, and $L_{xy} = D_{xy} - S_x \circ S_y$ where the symbol $\circ$ indicates element by element multiplication between two matrices, S_x and S_y are calculated using Eq. (4.3).

D_{xx} is a diagonal matrix of size $n \times n$ and its i^{th} diagonal element is the sum of the i^{th} row or column in $S_x \circ S_x$. Similarly, D_{yy} is a diagonal matrix of size $n \times n$ and its i^{th} diagonal element is the sum of the i^{th} row or column in $S_y \circ S_y$. Also, D_{xy} is a diagonal matrix of size $n \times n$ and its i^{th} diagonal element is the sum of the i^{th} row or column in $S_x \circ S_y$.

H_{xy}, H_{xx}, and H_{yy} are calculated in a similar manner to the way L_{xy}, L_{xx}, and L_{yy} are calculated where $H_{xx} = G_{xx} - B_x \circ B_x$, $H_{yy} = G_{yy} - B_y \circ B_y$, and $H_{xy} = G_{xy} - B_x \circ B_y$ where B_x can be calculated using Eq. (4.4). G_{xx} is a diagonal matrix of size $n \times n$ and its i^{th} diagonal element is the sum of the i^{th} row or column in $B_x \circ B_x$. Similarly, G_{yy} is a diagonal matrix of size $n \times n$ and its i^{th} diagonal element is the sum of the i^{th} row or column in $B_y \circ B_y$. G_{xy} is a diagonal matrix of size $n \times n$ and its i^{th} diagonal element is the sum of the i^{th} row or column in $B_x \circ B_y$.

The first term in the objective function $U_x H_{xy} U_y^T$ is the globality preserving term which preserves the global structure. Additionally, $XL_{xy}Y^T$ is the locality preserving term which preserves the local structure. To obtain the projection matrices W_x^T and W_y^T, the optimization problem Eq. (4.9) is converted to a generalized eigenvalue decomposition problem as follows:

$$\begin{pmatrix} & A_{xy} \\ A_{xy}^T & \end{pmatrix} \begin{pmatrix} W_x \\ W_y \end{pmatrix} = \lambda \begin{pmatrix} A_{xx} & \\ & A_{yy} \end{pmatrix} \begin{pmatrix} W_x \\ W_y \end{pmatrix}, \tag{4.10}$$

where $A_{xy} = U_x H_{xy} U_y^T + \beta X L_{xy} Y^T$, $A_{xx} = U_x H_{xx} U_x^T + \beta X L_{xx} X^T$, and $A_{yy} = U_y H_{yy} U_y^T + \beta Y L_{yy} Y^T$ are positive semi definite matrices. Let $W^T = [W_x W_y]^T$, $A = \begin{pmatrix} & A_{xy} \\ A_{xy}^T & \end{pmatrix}$, and $E = \begin{pmatrix} A_{xx} & \\ & A_{yy} \end{pmatrix}$ then Eq. (4.10) can be rewritten as:

$$AW = \lambda EW. \tag{4.11}$$

The matrix W is computed by finding the eigenvalues of $E^{-1}A$ or more generally Eq. (4.11). Once the projection matrix W is computed, the pairs W_x and W_y are computed. Then, projected features are derived easily as $W_x^T X$, and $W_y^T Y$. GLPCCA attempts to ensure both local and global structure of data samples from the two sets X and Y are preserved in the subspace.

4.4 The Proposed MGLPCCA

In order to deal with the situations where three or more sets are available, MGLPCCA is proposed; in the spirit of MCCA [14]. Given P sets of random variables $X_1, ...X_P$, with dimensions $m_1, ...m_P$, the objective of MGLPCCA is to maximize the correlation of data samples from different heterogeneous sets while maintaining the embedded structure of data samples locally and globally. Mathematically, MGLPCCA is formalized as follows:

$$\begin{aligned} \max_{W} \quad & Trace\Bigg(WAW\Bigg) \\ \text{subject to} \quad & WEW = P(I), \end{aligned} \tag{4.12}$$

where $W = \Big(W_1 \quad W_2 \quad \cdots \quad W_P\Big)$ and is the projection matrices of the P sets, $A = \begin{pmatrix} 0_{1,1} & A_{1,2} & \cdots & A_{1,P} \\ A_{2,1} & 0_{2,2} & \cdots & A_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ A_{P,1} & A_{P,2} & \cdots & 0_{P,P} \end{pmatrix}$ embodies the cross covariances of each two sets, and $E = \begin{pmatrix} A_{1,1} & 0_{1,2} & \cdots & 0_{1,P} \\ 0_{2,1} & A_{2,2} & \cdots & 0_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{P,1} & 0_{P,2} & \cdots & A_{P,P} \end{pmatrix}$ represents the auto covariances of each set. A_{ij} is of dimension $m_i \times m_j$ and is obtained as $A_{ij} = U_{x_i} H_{ij} U_{x_j}^T + \beta X_i L_{ij} x_j^T$, $L_{ij} = D_{x_i x_j} - S_{x_i} \circ S_{x_j}$ and $H_{ij} = G_{x_i x_j} - B_{x_i} \circ B_{x_j}$. Similarly, A_{ii}, H_{ii}, and L_{ii} are computed. B_{x_i} and S_{x_i} are obtained using Eq. (4.3) and Eq. (4.4). Eq. (4.12) is further written as follows:

$$\max_{W_1 \ldots W_P} \quad Trace\left(\begin{pmatrix} W_1 & W_2 & \cdots & W_P \end{pmatrix} \begin{pmatrix} 0_{1,1} & A_{1,2} & \cdots & A_{1,P} \\ A_{2,1} & 0_{2,2} & \cdots & A_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ A_{P,1} & A_{P,2} & \cdots & 0_{P,P} \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_P \end{pmatrix}\right) \tag{4.13}$$

$$\text{subject to} \quad \begin{pmatrix} W_1 & W_2 & \cdots & W_P \end{pmatrix} \begin{pmatrix} A_{1,1} & 0_{1,2} & \cdots & 0_{1,P} \\ 0_{2,1} & A_{2,2} & \cdots & 0_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{P,1} & 0_{P,2} & \cdots & A_{P,P} \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_P \end{pmatrix} = P(I).$$

The objective function in Eq. (4.13) considers maximizing the correlation while maintaining the embedded structure of data samples locally and globally between any two different sets. To solve this optimization, Eq. (4.13) is formalized as a generalized eigenvalue problem as follows:

$$\begin{pmatrix} 0_{1,1} & A_{1,2} & \cdots & A_{1,P} \\ A_{2,1} & 0_{2,2} & \cdots & A_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ A_{P,1} & A_{P,2} & \cdots & 0_{P,P} \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_P \end{pmatrix} = \lambda \begin{pmatrix} A_{1,1} & 0_{1,2} & \cdots & 0_{1,P} \\ 0_{2,1} & A_{2,2} & \cdots & 0_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{P,1} & 0_{P,2} & \cdots & A_{P,P} \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_P \end{pmatrix}, \tag{4.14}$$

where $(W_1^T, W_2^T, ...W_P^T)^T$ is the projection matrix which is associated with the top-k eigenvalues λ.

4.5 The Human Action Recognition Framework.

The proposed human action recognition framework as shown in Figure 4.1 consists of three stages: computation of descriptors, the proposed fusion techniques, and classification. At the core of the framework is the proposed MGLPCCA. In this section, each stage is presented in details.

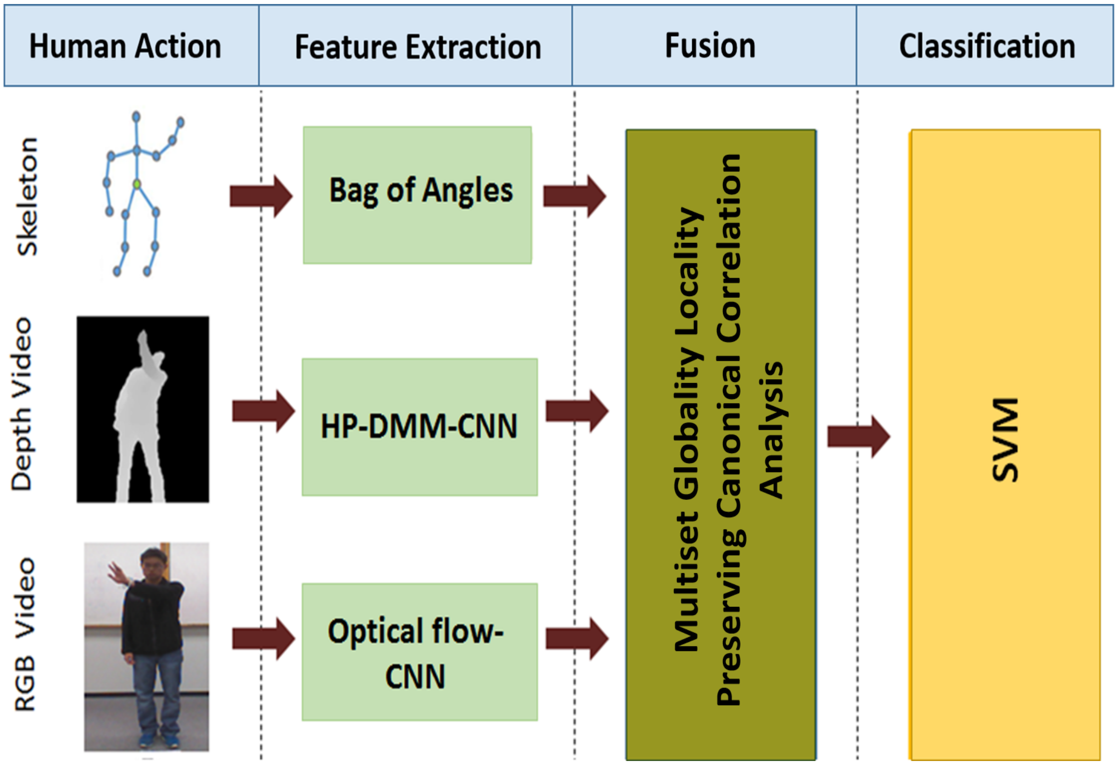


Figure 4.1: The proposed human action recognition framework using MGLPCCA considers three modalities including RGB video, depth video, and skeleton.

4.5.1 Descriptors

One of the main challenges facing human action recognition is feature extraction. Skeleton sequence, depth videos, and RGB videos data are rich sources of human movements and activities. For depth videos, (HP-DMM-CNN) is investigated. For RGB videos, optical flow frames using (TV-L1) is extracted at the beginning, then optical flow pre-trained ConvNet is used. For skeleton, BoA is used as a feature representation (Refer to Section 3.3.1).

Depth Video

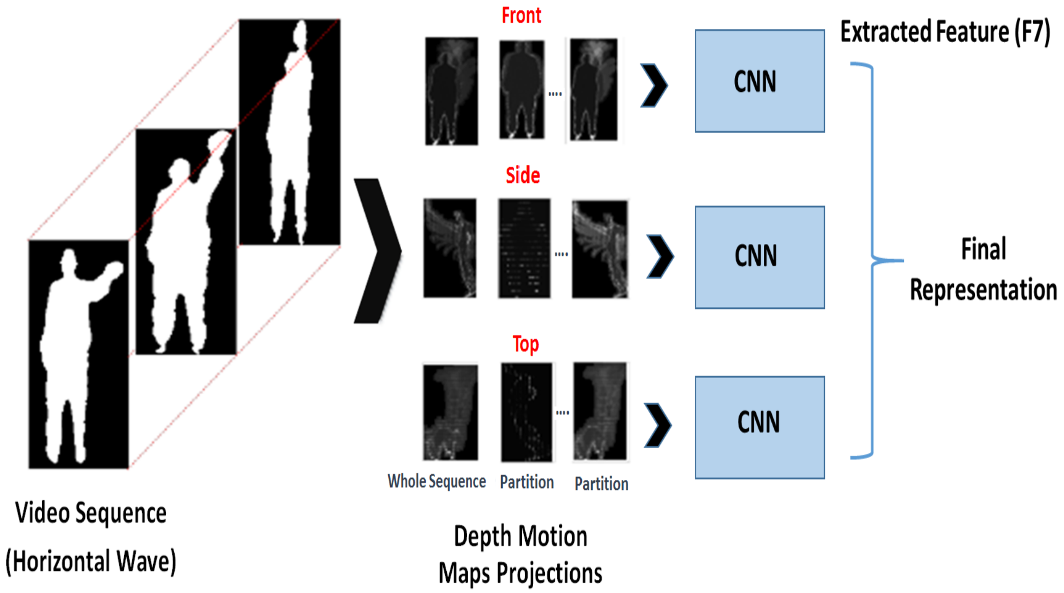


Figure 4.2: The pipeline of HP-DMM-CNN descriptor.

Depth videos are important in action recognition because of their ability to capture 3D action structure effectively [108]. Here, the Hierarchical Pyramid DMM Deep Convolutional Neural Network (HP-DMM-CNN) descriptor is proposed as a feature for depth videos as illustrated in Figure 4.2.

HP-DMM has the ability to capture more detailed information about the actions and fine changes of human movements, due to its ability to capture the sub-actions within a video sequence. To construct HP-DMM, the video sequence is divided into multiple parts or partitions of equal number of frames. Here, we employ a two-level pyramid with three partitions; which showed good results in [90]. The construction of a hierarchical pyramid of two levels is shown in Figure 3.5.

Finally, Deep ConvNet [109] is used to describe the local shape of each $DMM_{F,S,T}$ in each partition. The architecture of Deep ConvNet is identical to that in [109]. Assume $C(k, n, s)$ is a ConvNet with kernel size $k \times k$, n filters and stride s, $P(k, s)$ is a max pooling layer of kernel size $k \times k$ and stride s, Nor is a normalized layer, $ReLU$ is a Rectified Linear Unit (ReLU), $FC(n)$ is a fully connected layer with n filters and $D(r)$ is a dropout layer with dropout ratio r. The architecture of the network is as follows: $C(7, 96, 2) - ReLU - P(3, 2) - Nor - C(5, 384, 2) - ReLU - P(3, 2) - Nor - C(3, 512, 1) - ReLU - C(3, 512, 1) - ReLU - C(3, 384, 1) - ReLU - P(3, 2) - FC(4096) - D(0.5) - FC(4096) - D(0.5) - FC(|A| + 1)$. The output from the second fully connected layer is used as a descriptor for each $DMM_{F,S,T}$ in each partition. We used the publicly available $VGG - f$ pretrained on ImageNet ILSVRC-2012 challenge dataset [109].

RGB Video

An optical flow CNN is used to capture motion. The final descriptor is the fusion of all time descriptors. For each pair of successive frames, optical flow is computed as in [110]. Then, the optical flow pretrained ConvNet is used to extract the descriptor for each optical flow frame f_t. The architecture of Deep ConvNet is identical to that in [111]. The architecture of the network is as follows: $C(7, 96, 2) - ReLU - P(3, 2) - Nor - C(5, 384, 2) - ReLU - P(3, 2) - Nor - C(3, 512, 1) - ReLU - C(3, 512, 1) - ReLU - C(3, 384, 1) - ReLU - P(3, 2) - FC(4096) - D(0.5) - FC(4096) - D(0.5) - FC(|A| + 1)$. Here, we used the publicly available $VGG - f$ pretrained on UCF101 [111]. The output from the second fully connected layer is used as a descriptor for each optical flow frame. Finally, the descriptor Des is computed by minimum and maximum pooling as follows:

$$Des = [\min_{1<t\leq T}(f_t) \quad \min_{1<t\leq T}(f_t - f_{t-1}) \quad \max_{1<t\leq T}(f_t) \quad \max_{1<t\leq T}(f_t - f_{t-1})]^T, \tag{4.15}$$

where T is the number of frames per video and f_t is the output of second fully connected layer at time t.

4.5.2 Fusion and Classification

The shared subspace is learned using the proposed MGLPCCA. Following [112], the final descriptor is obtained as follows

$$F = \left(\alpha_1 W_1^T X_1 + \alpha_2 W_2^T X_2 + ... + \alpha_P W_P^T X_P\right), \tag{4.16}$$

where W_i is the projection matrix for the i^{th} set X_i and α_i is the weight for each projected feature. In the experiments, values set included $\alpha_i = 1/P$ as in [112] which essentially means average pooling over the projected features. This representation is adopted because of its simplicity and good performance. Finally, Linear SVM is adopted for classification [98].

4.6 Experimental Results

In order to evaluate the effectiveness of the proposed framework, comprehensive experiments are conducted on several publicly available multiset human action datasets. In this section, dataset description, experimental settings, and performance evaluations are presented. First the datasets and experimental settings are introduced. Then, analytical evaluations are presented to justify the improvement introduced by the HP-DMM-CNN. Then, the fused feature representation in MGLPCCA subspace are visualized. Finally, the recognition performance on MGLPCCA against some state-of-the-art methods are compared.

4.6.1 Multiset Action Datasets

We conducted the experiments on five different action datasets: MSR Action3D [94], UTD Multimodal Human Action Dataset (UTD-MHAD) [95], Multimodal Action Database (MAD) [113], Kinect Activity Recognition Database [114], and SBU Kinect interaction dataset [115].

The **MSR Action3D Dataset** contains 1114 action sequences (557 depth sequence, and 557 skeleton sequences) for 20 actions performed by 10 subjects. All the sequences were captured with Kinect sensor. The 20 actions are: *high wave, horizontal wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pickup throw.* Each subject performed every action 2 or 3 times. We followed the same settings in [78], where all the 20 actions were employed. Half of the subjects were used for training and the rest for testing.

The **UTD Multimodal Human Action Dataset** (UTD-MHAD) is composed of 3444 sequences (861 RGB video sequences, 861 depth video sequences, 861 skeleton sequences, and 861 accelerometer sensor data) for 27 actions performed by 8 subjects. RGB sequences, depth sequences, and skeleton sequences were captured using Kinect sensor. Accelerometer data sequences were captured by a wearable inertial sensor. The 27 actions are: *right arm swipe to the left, right arm swipe to the right, right hand wave, two hand front clap, right arm throw, cross arms in the chest, basketball shoot, right hand draw x, right hand draw circle, right hand draw circle, draw triangle, bowling, front boxing, baseball swing from right, tennis right hand forehand swing, arm curl, tennis serve, two hand push, right hand knock on door, right hand catch an object, right hand pick up and throw, jogging in place, walking in place, sit to stand, stand to sit, forward lunge, and squat.* Every subject performed each action 4 times. We followed the experimental settings in [95], where the 27 actions were employed. Half of subjects were used for training and the other half for testing.

The **Multimodal Action Database** (MAD) consists of 120 sequences (40 depth video sequences, 40 RGB video sequences, and 40 skeleton sequences) for 35 actions performed by 20 subjects. Depth sequences and RGB video sequences were both captured using Kinect sensor. The 35 actions are: *running, crouching, jumping, walking, jump and side-kick, left arm swipe to the left, left arm swipe to the right, left arm wave, left arm punch, left arm dribble, left arm pointing to the ceiling, left arm throw, swing from left (baseball swing), left arm receive, left arm back receive, left leg kick to the front, left leg kick to the left, right arm swipe to the left, right arm swipe to the right, right arm wave, right arm punch, right arm dribble, right arm pointing to the ceiling, right arm throw, Swing from right (baseball swing), right arm receive, right arm back receive, right leg kick to the front, right leg kick to the right, cross arms in the chest, basketball shooting, both arms pointing to the screen, both arms pointing to both sides, both arms pointing to right side, and both arms pointing to left side.* Every subject performed each action 2 times. Half of subjects were used for training, and the others for testing.

The **Kinect Activity Recognition Dataset** (KARD) is composed of 18 actions performed by 10 subjects. Every subject performed each action 3 times. The dataset was captured using kinect camera. The 18 actions are: *Horizontal arm wave, high arm wave, two hand wave, catch cap, high throw, draw X, draw tick, toss paper, forward kick, side kick, take umbrella, bend, hand clap, walk, phone call, drink, sit down, and stand up.* In our experiments, two modalities are used which are RGB camera, and depth video sequences, followed the same settings in [114], where all the 20 actions were employed. Half of the subjects were used for training and the rest of subjects were used for testing.

The **SBU Kinect Interaction Dataset** (SBU) was presented to recognize the interaction between two persons. Depth video sequences and RGB video sequences are captured by kinect camera. SBU is composed of 8 actions performed by 21 subjects.

Note that in most interactions, one person is acting and the other person is reacting. The dataset was captured using kinect camera. The 8 actions are: *approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands.* In the experiments, two modalities are used which are RGB camera, and depth video sequences, following the same settings in [115], where 5-fold cross-validation scheme is employed.

4.6.2 Analytical Evaluation of the Proposed Descriptors

In order to verify the performance achieved by the proposed descriptor, experiments are conducted on the Action3D dataset, by analyzing the performance of the HP-DMM-CNNs against other depth descriptors.

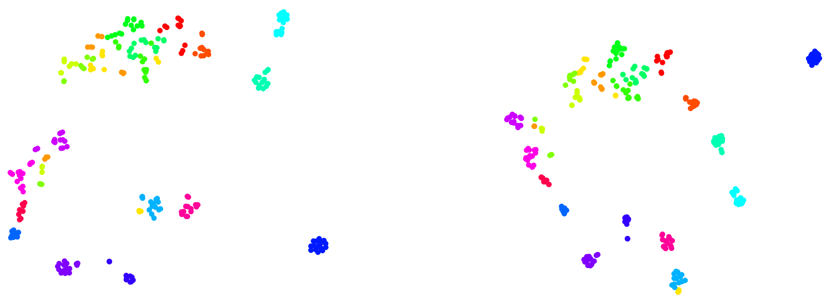
Evaluation of HP-DMM-CNN The performance of the proposed HP-DMM-CNN is compared against different baseline depth descriptors. As shown in Table 4.1, the recognition accuracy of the HP-DMM-CNN outperforms other baseline descriptors. The evaluation results indicate that the proposed HP-DMM-CNN is able to model the temporal dynamics, mitigate the speed variations of actions, and gain the discriminatory power of CNNs as a feature extractor.

Table 4.1: Recognition Accuracy comparison between HP-DMM-CNN against other baseline depth descriptors on MSR Action3D Dataset.

Depth Descriptor	Recognition Accuracy %
DMM-HOG [8]	84.2
HOG3D-SC [116]	87.1
HON4D [65]	88.9
ROP [9]	86.5
HP-DMM-CNN	**92.31**

4.6.3 Qualitative Analysis of MGLPCCA

To gain qualitative insight into feature representations learned from MGLPCCA, the visualization of the subspaces for CCA and MGLPCCA on the MSR Action3D dataset using t-SNE [97] is constructed. Figure 4.3 visually illustrates the fused feature representations in CCA, and MGLPCCA subspaces, respectively. From the figure, it is observed that the between class distance is small and the classes are smeared in CCA subspace. Conversely, similar classes are more properly separated in MGLPCCA subspace, where the samples in the same class are more compactly clustered, leading to a more discriminative subspace.



(a) t-SNE embedding of CCA.

(b) t-SNE embedding of MGLPCCA.

Figure 4.3: The visualization of the fused feature representation in (a) CCA and (b) MGLPCCA subspaces. Classes are more compact and less smeared in MGLPCCA than in CCA.

4.6.4 Recognition Performance Evaluation

The experimental results are reported on the five datasets, MSR Action3D, UTD Multimodal Human Action Dataset (UTD-MHAD), Multimodal Action Database (MAD), Kinect Action Recognition Dataset (KARD), and SBU Kinect Interaction Dataset. The effectiveness of MGLPCCA is verified by conducting the same experiments on the baseline techniques (CCA [13], MCCA [14], Locality Preserving Canonical Correlation Analysis (LPCCA) [107], and Multiset Locality Preserving Canonical Correlation Analysis (MLPCCA) [117]) and reporting the results in terms of the recognition accuracy. Since the features of optical flow and depth videos are extracted from the second fully connected layer of the Deep ConvNet, the dimensionality of the extracted features is extremely high. In order to overcome this problem, Principal Component Analysis (PCA) is adopted for dimensionality reductions to only one hundred dimensions.

MSR Action3D dataset

First, the recognition accuracy for each set is studied. As the dataset MSR Action3D has only depth and skeleton, recognition accuracy experiments on only HP-DMM-CNN and BoA are conducted and the results tabulated in Table 4.2. These results show that HP-DMM-CNN and BoA have a recognition accuracies of 92.31 % and 86.92 %, respectively.

Table 4.2: Recognition Accuracy for each feature set on MSR Action3D dataset.

Feature Set	Recognition Accuracy %
HP-DMM-CNN	92.31
BoA	86.92

Table 4.3 illustrates the recognition accuracy of CCA and MGLPCCA by fusing HP-DMM-CNN and BoA descriptors. It is observed that MGLPCCA outperforms CCA and LPCCA by around 2.5 % and 1.2 %, respectively, due to preserving both the global and local structure.

Table 4.3: Recognition Accuracy Comparison among MGLPCCA,LPCCA, and CCA on MSR Action3D Dataset.

Fusion Technique	Recognition Accuracy %
CCA	94.23
LPCCA	95.77
Proposed MGLPCCA	**96.92**

Table 4.4: Recognition Accuracy for each feature set on UTD-MHAD Dataset.

Feature Set	Recognition Accuracy %
HP-DMM-CNN	82.79
BoA	85.35
Optical flow CNN	82.56

UTD-MHAD dataset

First, the recognition accuracy for each set with the UTD-MHAD dataset is examined. The recognition performances of each feature set, HP-DMM-CNN, BoA, and optical flow CNN, are first computed and recorded in Table 4.4. The results show that the BoA has the highest recognition accuracy with 85.35 %.

To realize the benefits of fusion using MGLPCCA for two or three sets, experiments on all the possible combinations of the sets (skeleton, depth, optical flow) are conducted and presented in Table 4.5. It is noted from the table that MGLPCCA has higher recognition accuracy than CCA and LPCCA. The best combination of sets in terms of recognition accuracy is the *Optical flow-Skeleton* combination. Furthermore, from the Table 4.5, we observe that the fusion of three sets using MGLPCCA results in the highest recognition accuracy, 95.35 %. This demonstrates that the fusion of three sets using MGLPCCA leads to superior performance versus the fusion of any two sets.

Table 4.5: Recognition Accuracy Comparison Among MGLPCCA, LPCCA, MLPCCA, CCA, and MCCA on UTD-MHAD Dataset.

Fusion Technique	Optical flow-Depth %	Depth-Skeleton %	Optical flow-Skeleton %	Optical flow-Depth-Skeleton %
CCA	88.84	90.7	93.02	-
MCCA	-	-	-	92.79
LPCCA	90.93	91.16	92.09	-
MLPCCA	-	-	-	93.95
Proposed MGLPCCA	**93.26**	**92.56**	**93.49**	**95.35**

Table 4.6: Recognition Accuracy for each feature set on MAD Dataset.

Feature Set	Recognition Accuracy %
HP-DMM-CNN	61.78 %
BoA	85.5 %
Optical flow CNN	77.12 %

MAD dataset

First, we look into the recognition accuracy by each of the three feature sets, HP-DMM-CNN, BoA, and optical flow CNN, in MAD dataset. As shown in Table 4.6, BoA shows the highest recognition accuracy with 85.5 %. To realize the benefits of fusion using MGLPCCA for two or three sets, experiments on all the possible combinations of skeleton, depth, and optical flow are conducted and presented in Table 4.7. The first thing we notice from Table 4.7 is that the fusion of any two or three sets outperforms any single set. The proposed MGLPCCA yields recognition accuracy of 91.07 %, 93.77 %, 95.12 % for the combinations of *Optical flow-Depth, Depth-Skeleton, and Optical flow-Skeleton*, respectively, beating CCA. Finally, MGLPCCA leads the way with the best recognition accuracy of 96.63 % among the rest.

Table 4.7: Recognition Accuracy Comparison Among MGLPCCA, LPCCA, MLPCCA, CCA, and MCCA on MAD Dataset.

Fusion Technique	Optical flow-Depth %	Depth-Skeleton %	Optical flow-Skeleton %	Optical flow-Depth-Skeleton %
CCA	86.70	92.09	94.44	-
MCCA	-	-	-	95.79
LPCCA	80.64	93.26	87.71	-
MLPCCA	-	-	-	95.62
Proposed MGLPCCA	**91.07**	**93.77**	**95.12**	**96.63**

KARD dataset

Again, the recognition accuracy by an individual feature set is studied, HP-DMM-CNN and optical flow CNN, and the results are summarized in in Table 4.8. This table shows that optical flow CNN has the higher recognition accuracy, 92.96 %. To reveal the benefits of fusion using MLPCCA for two sets, experiments on the fusion of depth and optical flow are conducted and the results are presented in Table 4.9. It is noted from Table 4.9 that MGLPCCA has higher recognition accuracy than CCA and LPCCA. MGLPCCA yields a perfect recognition accuracy of 100 %.

Table 4.8: Recognition Accuracy for each set on KARD Dataset.

Feature Set	Recognition Accuracy %
HP-DMM-CNN	87.78
Optical flow CNN	92.96

Table 4.9: Recognition Accuracy Comparison between the Proposed Fusion Technique MGLPCCA, LPCCA and CCA on KARD Dataset.

Fusion Technique	Optical flow-Depth %
CCA	98.52
LPCCA	97.40
Proposed MGLPCCA	**100**

SBU Kinect Interaction dataset

Once more, the recognition accuracy of each set is computed for HP-DMM-CNN, and optical flow CNN, respectively. Table 4.10 reports the recognition results which show that HP-DMM-CNN achieves the highest recognition accuracy, 84.48 %. Then, depth and optical flow are fused together using MGLPCCA, CCA and LPCCA. The comparison results are presented in Table 4.11. It is observed from Table that MGLPCCA provides the highest recognition accuracy of 90.1 %.

Table 4.10: Recognition Accuracy for each Feature set on SBU Kinect Interaction Dataset.

Feature Set	Recognition Accuracy %
HP-DMM-CNN	84.48
Optical flow CNN	82.81

Table 4.11: Recognition Accuracy Comparison between the Proposed Fusion Technique MGLPCCA, LPCCA, and CCA on SBU Kinect Interaction Dataset.

Fusion Technique	Optical flow-Depth %
CCA	88.43
LPCCA	87.9
Proposed MGLPCCA	**90.1**

4.7 Conclusion

In this chapter, Multiset Globality Locality Preserving Canonical Correlation Analysis (MGLPCCA) is proposed for learning the common subspace from two or more sets. In feature extraction, HP-DMM-CNN for depth is presented. Analytical evaluations unveiled the superiority of HP-DMM-CNN over several baselines. Moreover, the fused MGLPCCA subspace is assessed visually using t-SNE embedding, showing the discriminative ability of MGLPCCA over CCA. Based on MGLPCCA, a human action recognition framework with multiset fusion is introduced. The experimental results showed the effectiveness of the proposed framework; attributed to the ability of MGLPCCA in preserving the global and local structure of the data samples with the same time complexity of CCA.

Chapter 5

Action Recognition via Deep Learning

The work presented in the previous two chapters improved action recognition in datasets with a comparably small number of samples. However, for larger datsets, the effectiveness of such frameworks is limited. For example, it is quite expensive to compute just the covariance matrix, required for fusion methods. For a whole dataset containing thousands of videos, given that it has a time complexity of $O(Nd^2)$, where N is the number of samples in a given datasets and d is the dimension of the features, this becomes quickly intractable. Therefore in this chapter, we address the aforementioned drawbacks by migrating from the fusion of handcrafted and deep learning features to an end-to-end deep learning framework. The problem of video based action recognition via deep learning is explored for larger datasets compared to the datasets used in the two previous chapters. Here, action recognition performance is improved by finding an effective temporal and spatial representation. For capturing the temporal representation, two methods are introduced for improving long term temporal information modelling including Temporal Relational Network (TRN) and Temporal Second Order

Pooling based Network (T-SOPN). Moreover, the representation is harnessed by using complementary learning techniques including Global-Local Network (GLN) and Fuse-Inception Network (FIN). Additionally, a two stream network fusing RGB and optical flow, is adopted for improving the final performance.

5.1 Introduction

With the success achieved by deep learning in addressing different computer vision challenges, researchers put forth effort to address video based action recognition using deep learning. The initial attempts were based on Deep Convolutional Neural Networks (Deep ConvNets) architectures [36]. These architectures were chosen due to its magnificent capabilities of generalization and capturing discriminative features. However, at the beginning, video based action recognition using 2D Deep ConvNets did not obtain the level of success gained in image recognition using 2D Deep ConvNets. The reason is that the nature of videos is different from that of images. In other words, video based action recognition is hindered by more challenges than image recognition. For instance, [36] and [37] focus only on short term temporal information, limiting the performance. Although some researchers proposed 3D Deep ConvNets to capture the temporal dynamics needed for better representation (i.e 64 frames) [16], such networks utilize immense large memory which limits the temporal duration to modeling. Wang *et al.* [118] introduced good practices dealing with the temporal dynamics for this endeavour and won the Activity-Net Challenge 2016 showing the effectiveness of capturing long term temporal dynamics.

Moreover, huge improvement can be achieved from representing the scene of the action by integrating complementary information about the action itself. In [119], complementary information improves the recognition accuracy. Additionally, the performance can be boosted by capturing global and local information.

The above challenges motivated the study herein of video based action recognition by improving the long term temporal dynamics modelling which is called *Temporal Learning*. Moreover, it also motivated leveraging complementary information as a method of improving the discriminative ability of the learned representation which is called *Complementary Learning*. Figure 5.1 illustrates the proposed temporal and complementary learning framework for video.

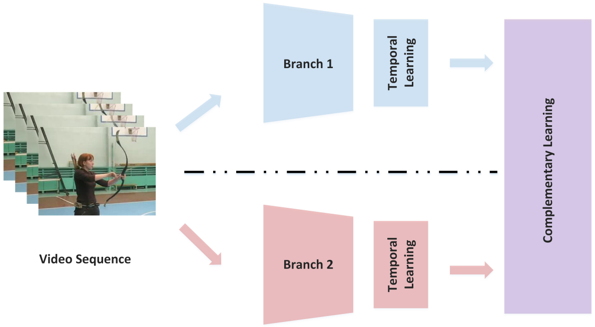


Figure 5.1: The proposed temporal and complementary learning framework. The input video sequence (RGB or Optical flow) is forwarded to two branches. At each branch, the long term temporal information is modelled by temporal learning. Moreover, the learned representation of each branch is harnessed by complementary learning.

To improve long term temporal dynamics, two networks are proposed to better leverage the long term temporal modelling. First, TRN is introduced, which combines 2D Deep ConvNets and CRF in an end-to-end training framework. In TRN, each video is divided into a fixed number of partitions and CRF layer models the long term information embedded in the video. Second, T-SOPN is presented to model the temporal structure of the video (as inspired by[120]). T-SOPN models the long term temporal dynamic cues through covariance operation.

Moreover, to improve the overall recognition performance of the video, two effective frameworks are proposed based on the concept of complementary learning. First, GLN which combines both global information and the local discriminative features, corroborating the discriminative representation, is presented. Second, inspired by [121], the idea of Dual Networks was modified and FIN was introduced to leverage the improved discriminative ability of the learned feature representation through fusion and complementary learning. The former attempts to learn a distinct representations describing the action. The latter fuse two representations for a better performance.

5.2 Temporal Learning

For the sake of ameliorating the long term temporal modelling, TRN and T-SOPN are introduced.

5.2.1 Temporal Relational Network (TRN)

Relationships among sub-actions within the action play a great role in representing the action itself. The proposed TRN exploits the statistical relationship among the sub-actions within the whole action video as shown in Figure 5.2. Understanding the action requires modelling the interactions between partitions within the video. The proposed TRN utilizes 2D Deep ConvNet and CRF to learn the temporal long term relationships among the partitions. CRF [122] is a class of undirected graph models which consolidates statistical relations into a discriminative task.

Assume that video V is divided into N partitions. Each partition S_i is representing the i^{th} sub-action/partition. Specifically: CRF can be formulated as follows:

$$p(r_1, r_2 ... r_N | f_1, f_2, ..f_N) = \frac{1}{Z} exp(\Phi(r_1, r_2 ... r_N | f_1, f_2, ..f_N; W)), \tag{5.1}$$

where r_i and f_i represent the class and feature of the i^{th} partition, respectively. The variable W represents the model parameters, the variable Z denotes the normalizing constant, and the variable Φ is the joint potential function and is expressed as follows:

$$p(r_i|(r_l)_{l=1:N,l\neq i}, f_i; W) \propto \quad exp(\psi(r_i|f_i; W_i)) + \Sigma_{i,j=1,i\neq j}^{N}\varphi_{ij}(r_i, r_j|W_{i,j}), \tag{5.2}$$

where the unary potential ψ associates the i^{th} partition with the feature representation f_i. The variable φ is the binary potential and captures the statistical relationship among the N partitions. CRF has been adopted as a solution to many computer vision challenges since it captures the statistical relationships. However, there is the issue of the intractability of computing the normalizing constant Z, especially when cycles are present in the graph. Inspired by [123], the problem is reformulated for temporal long term modelling. The posterior distribution of the i^{th} partition of r_i is computed as follows:

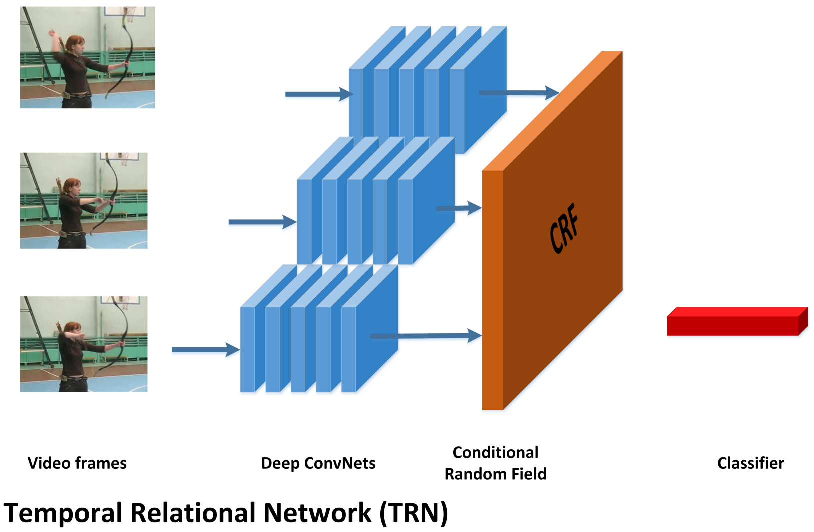


Figure 5.2: The proposed TRN. In TRN, the video is divided into partitions where the temporal information is modelled using CRF.

$$p(r_i|r_1, r_2, ..r_N, f_i; W) \propto \quad exp(\psi(r_i|f_i; W_i)) + \Sigma_{i,j=1,i\neq j}^{N}\varphi_{ij}(r_i, r_j|W_{i,j}). \tag{5.3}$$

According to [122], $\psi(r_i|f_i)$ is usually chosen to be a linear function of f_i for each r_i. So, q_i can be explicitly expressed as follows:

$$q_i = \sigma(W_i f_i + \Sigma_{i,j=1,i\neq j}^{N} W_{i,j} 1_j), \tag{5.4}$$

where σ represents the softmax function. 1_j is the one hot vector for the j^{th} partition. Eq. (5.4) is generalized for the case of r_j is not deterministic but is given through the posterior probability q_j as follows:

$$q_i = \sigma(W_i f_i + \Sigma_{i,j=1,i\neq j}^{N} W_{i,j} q_j). \tag{5.5}$$

In the training phase, q_i^t is the posterior probability at time t for the i^{th} partition and the posterior probability at time $t+1$ is denoted as follows:

$$q_i^{t+1} = \sigma(W_i f_i + \Sigma_{i,j=1,i\neq j}^{N} W_{i,j} q_j^t). \tag{5.6}$$

The above formula is considered as a building block for the Temporal Relational Network and is used for iterative training to capture the inter-dependencies among the sub-actions/partitions. TRN can also be considered as a special form of the Recurrent Neural Network (RNN) – at each step it takes in a fixed set of inputs, i.e. the observed features f_1, f_2, $..f_N$ and refines the estimates of posterior probabilities.

5.2.2 Temporal Second Order Pooling Based Network (T-SOPN)

Temporal Second Order Pooling (TSOP), which is the pillar of T-SOPN, is the temporal version of Statistically Motivated Second Order (SMSO) pooling [120]. The second order pooling techniques produce tremendously large representations compared to the first order pooling techniques [124]. The goal of SMSO pooling is to produce an effective, yet compressed second order pooled representation. The video V is divided into N partitions. The extracted features of each partition are applied to TSOP, capturing an effective whole video representation. Figure 5.3 shows a network with TSOP pooling. Let $X \in \Re^{n \times c}$ denotes a data matrix with n samples and c channels. The variable X represents the extracted features from the N partitions and are concatenated across the depth channel $c = \Sigma_{i=1}^{N} c_i$. Effectively, the second order pooling is the covariance matrix $Y_{cov} \in \Re^{c \times c}$ and can be computed as follows:

$$Y_{cov} = \frac{1}{n-1} \Sigma_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T = \frac{1}{n-1} \tilde{X}^T \tilde{X}, \tag{5.7}$$

where $\tilde{X}$ is the mean subtracted data matrix, $x_i \in \Re^c$ is the i^{th} sample and μ is the mean of the data samples.

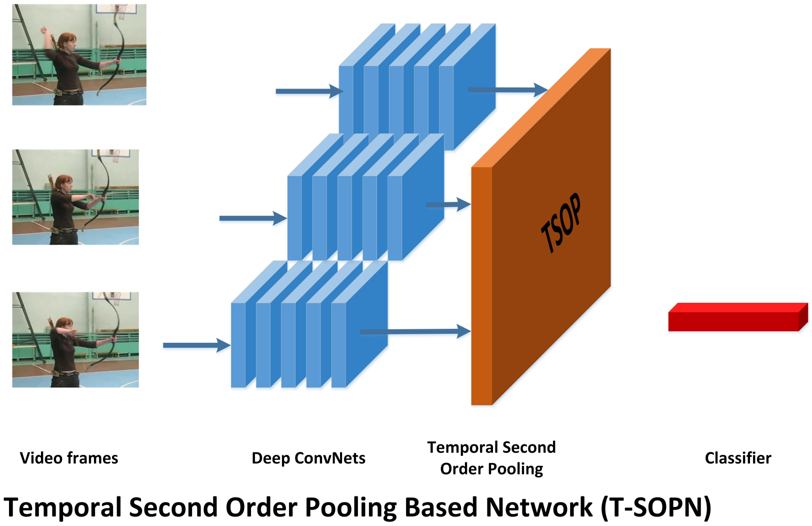

TSOP
Video frames
Deep ConvNets
Temporal Second Order Pooling
Classifier
Temporal Second Order Pooling Based Network (T-SOPN)

Figure 5.3: The proposed Temporal Second Order Pooling based Network (T-SOPN) where TSOP is used to pool features from the partitioned video.

SMSO pooling yields a compressed second order representation during parametric vectorization. It relies on a parametric vectorization layer with trainable weights $W \in \Re^{c \times p}$, with p representing the dimensionality of resulting vector. Assume z as an output vector of the parametric vectorization layer and z_j is the j^{th} dimension which is computed as follows:

$$\frac{1}{\sqrt{\alpha}} z_j = \sqrt{w_j^T Y_{cov} w_j}, \tag{5.8}$$

$$= \sqrt{w_j^T \Sigma_{i=1}^n (x_i - \mu)(x_i - \mu)^T w_j}, \tag{5.9}$$

$$= \sqrt{\Sigma_{i=1}^n (w_j^T \tilde{x_i})^2}. \tag{5.10}$$

z is computed by a convolutional layer with size of $(1, 1, c, p)$ without bias. The output is applied to l_2 pooling and scaled by the constant $\sqrt{\alpha}$. The network is trained in an end-to-end manner where the temporal pooling is applied directly after the last convolutional layer. This version is named TSOP-V1.

The above equation can be simplified slightly to just one convolutional layer with a size of $(1, 1, c, p)$, TSOP-V2. In the experimental results section, TSOP-V1, TSOP-V2, and average pooling are studied as candidates for temporal pooling.

5.3 Complementary Learning

For the sake of improving the appearance modelling, two networks: Global-Local Network (GLN) and Fuse-Inception Network (FIN) are introduced.

5.3.1 Global-Local Network (GLN)

2D Deep ConvNets consist of a stack of convolutional layers interleaved with non-linear and pooling layers. Each convolutional layer is composed of a set of parameterized kernels capturing the local patterns along the input channels. Early layers of the network are more sensitive to local stimuli. In other words, earlier layers capture the spatial local information, while the later and intermediate layers capture a more global intuition about the input image. Moreover, the later layers are more invariant to transformations including translation, occlusion, and truncation of the local stimulus. Recently, Hu *et al.* introduced Squeeze and Excitation Networks [125] to exploit the interchannel interdependencies. Inspired by [125] [126], squeeze and excitation block, the building block of Squeeze and Excitation Networks, overshadows the more informative local activations. In other words, squeeze and excitation block emphasizes important local features, lending the representation more discriminative power. This observation led to the idea of a combined architecture to learn the global information about the scene of action and the local information regarding the action itself resulting in a better effective representation.

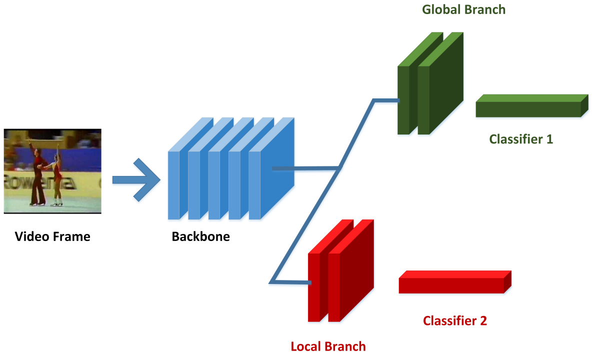


Figure 5.4: The proposed Global-Local Network (GLN). It is a two headed network consisting of the backbone and two branches (global and local branches).

Figure 5.4 shows the architecture composed of three main components: backbone, local branch, and global branch. The backbone is a fully convolutional acting as a feature extractor and is the common part of the whole architecture. It produces an intermediate level features capturing the details of the input video frames. The output activations are fed into the following parallel branches. Both branches have the same number of convolutional layers. For example, in the case of Inception V1, the backbone is the first 7 inception modules, and the last two inception modules $(5a, 5b)$ which are chosen as the basic architecture for the two branches. The global branch is the last two inception modules. Whereas, the local branch is based on the last two inception modules and modified squeeze and excitation block. The modified squeeze and excitation block is shown in Figure 5.5.

Suppose $X \in \Re^{H \times W \times C}$ is an input to the proposed modified squeeze and excitation block. First, two channel descriptors are computed by applying AVG and MAX pooling to capture the channel-wise information. The descriptor of MAX pooling with a kernel size (H, W) is applied to X. The c^{th} element of the descriptor Y^{max} is computed as follows:

$$Y_c^{max} = \max_{i=1:H, j=1:W} \quad X(i,j), c = 1, 2...C. \tag{5.11}$$

Moreover, the descriptor of average Pooling with kernel size (H, W) is applied to X. The c^{th} element of the descriptor Y^{avg} is computed as follows:

$$Y_c^{avg} = \frac{1}{H \times W} \sum_{i,j=1}^{H,W} \quad X(i,j), c = 1, 2...C. \tag{5.12}$$

Pooling results in vectors, which are highly invariant to translation and occlusion. In other words, they are a good way of representing the intermediate layers. To satisfy the properties desired from above, a gating mechanism is adopted by two fully connected

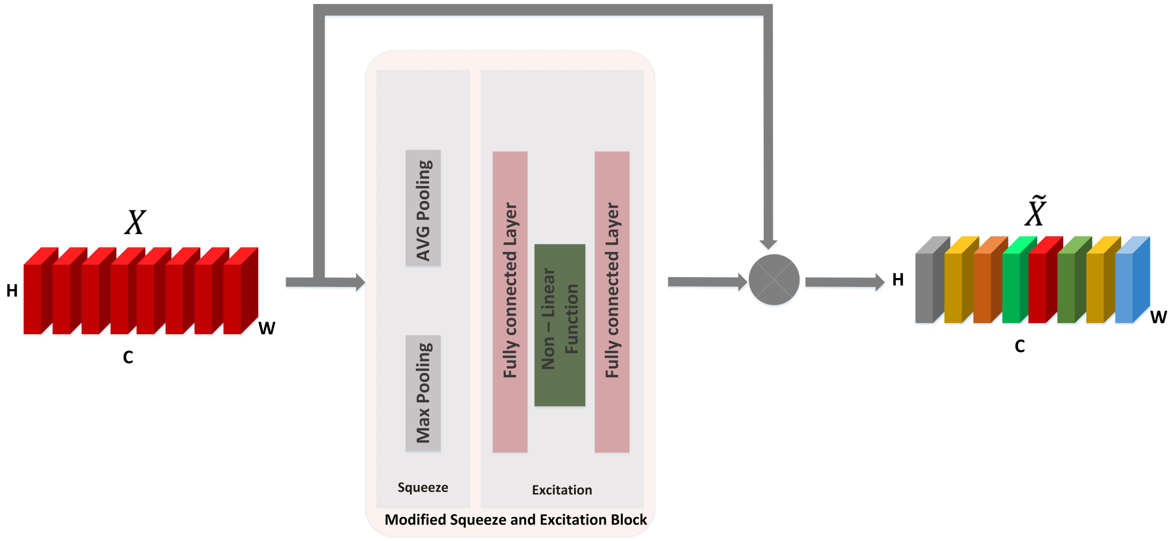


Figure 5.5: The modified squeeze and excitation block where the local feature maps are highlighted according their local information power.

layers around a non-linear layer which is mathematically formulated as follows:

$$S = Sigmoid(W_2 ReLU(W_1 Y)), \tag{5.13}$$

where $W_1 \in \Re^{C' \times C}$, and $W_2 \in \Re^{C \times C'}$ and C' is less than C. The variable S is the channels attention scale which is S^{max} for MAX pooling and S^{avg} for average pooling. The input is first dimensionally reduced to C', passed through a ReLU layer. Finally, the dimension is retained by W_2. The variable Y is either Y^{max} or Y^{avg}. In the experimental results section, one fully connected layer for both Y^{max} and Y^{avg} (Shared) and one fully connected layer for both Y^{max} and Y^{avg} (Separate) are studied.

The final output $\tilde{X}$ is computed by scaling X with the learned channels attention scale as in the following formula:

$$\tilde{X} = S^{max} \odot X + S^{avg} \odot X. \tag{5.14}$$

In other words, each activation channel is scaled by the weight from S^{max} or S^{avg}. If the MAX pooling is removed, the block will be downgraded to the squeeze and excitation block as in [125]. By applying the modified squeeze and excitation block, the network is lent a highlighting tool for the important local features.

Two possible architectures shown in Figure 5.6 are proposed. Architecture 1 is based on training two separate branches, each has its own loss. On the other hand, architecture 2 has two branches: the global branch and the fusion branch which sums the global and local branches. In the experimental results section, both architectures are studied.

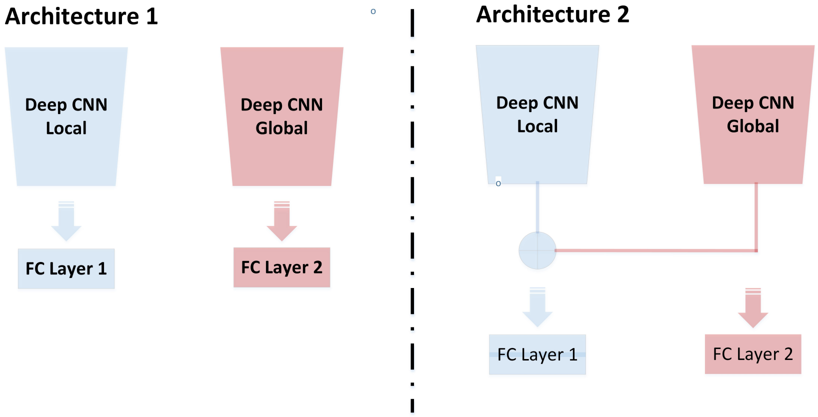


Figure 5.6: Possible architectures for Global-Local Network (GLN). In Architecture 1, the "global" and "local" branches are optimized. On the other hand, in Architecture 2, the global and "global + local" branches are optimized.

In order to train the network, the whole network is first trained on the global branch only. Then, the whole network is fine-tuned with the following loss function:

$$L = \lambda_G L_{global} + \lambda_L L_{Local}, \tag{5.15}$$

where L_{global} and L_{local} are the cross-entropy loss of global and local branches. λ_G and λ_L are the loss weights. Empirically, λ_G and λ_L are set to 0.1 and 1 respectively.

5.3.2 Fuse-Inception Network (FIN)

Inspired by [121], Fuse-Inception Network (FIN) is introduced. FIN discovers the representation to improve the recognition accuracy. FIN embodies two twin networks which are trained to fuse the intermediate feature as demonstrated in Figure 5.7. Our proposed network is based on Inception network [51], the first 7 inception modules are regarded as a backbone for both networks having the same structure. The feature representations are fused to the final representation. The idea of having an auxiliary classifier is inspired from [121] [60] which allows transfer learning from the auxiliary branch to improve the discriminative ability of the network. Moreover, complementarity is maintained by weighting the loss. The key idea is learning complementary presentation from video frames and enriching the hidden representation.

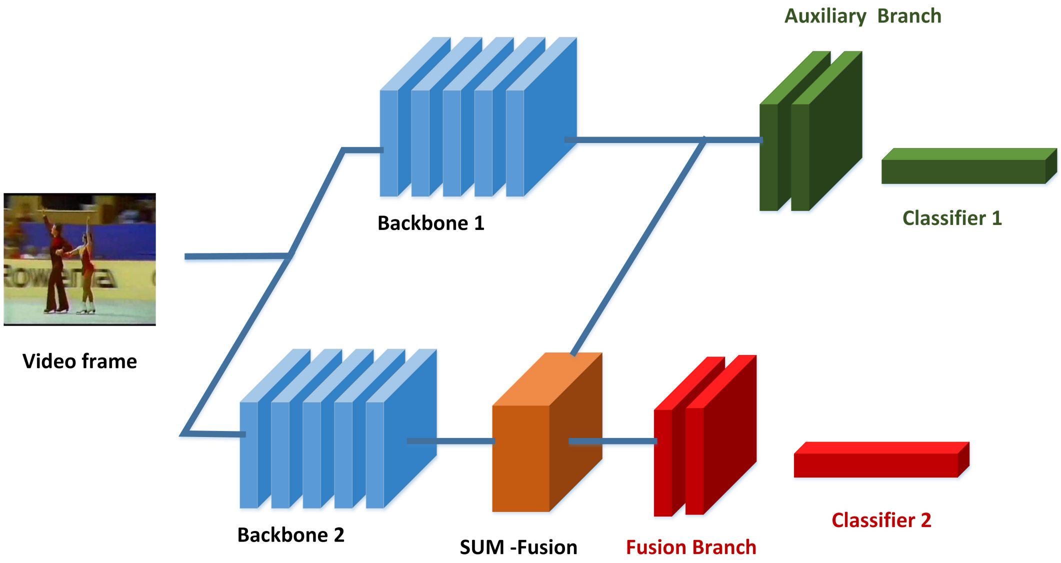


Figure 5.7: Fuse-Inception Network (FIN). FIN has two branches which are trained to capture complementary information regarding the action itself.

The core problem of FIN is the training strategy. Inspired by the training procedure of [121], a FIN training strategy is introduced which plays a vital role in learning the complementary features allowing for further performance improvement. The training procedure starts with training Inception network on the dataset. Then, the obtained weights are utilized for the upper branch (Backbone 1 and the auxiliary classifier) weight initialization. Then, the whole network is fine-tuned with the following loss function:

$$L = \lambda_F L_{fus} + \lambda_A L_{aux}, \tag{5.16}$$

where L_{fus} and L_{aux} are the fusion and auxiliary losses. The variables λ_F and λ_A are the loss weights for the fusion and auxiliary losses, respectively. The loss weights are chosen to be 1 and 0.1, respectively. The auxiliary loss weights play a vital role in learning the complementary features. Moreover, they act as a regularizer allowing for more importance to the L_{fus}. The final predication is computed as follows:

$$P = \lambda_F P_{fus} + \lambda_A P_{aux}, \tag{5.17}$$

where P_{fus} and P_{aux} are the prediction of the fusion and auxiliary networks. The variables λ_F and λ_A are the prediction weights.

5.4 Experimental Results

In this section, the datasets and the implementation details are presented. Then, the detailed analysis of the proposed networks is discussed. Finally, a comparison is made between the performance of the proposed framework against some of state-of-the-art techniques.

5.4.1 Datasets

Experiments were conducted on two open datasets of trimmed videos namely UCF101 [127] and HMDB-51 [128] datasets. The former embodies 101 action classes and 13320 videos. The same experimental settings were followed in [127] where the dataset is split into three training/testing splits for evaluation. The HMDB-51 dataset contains large diverse videos collected from different sources including web videos and movies. The dataset embodies 51 action classes and 6766 videos. As in [128], the dataset is divided into three training/testing splits for evaluation.

5.4.2 Implementation Settings

Following the guideline set in [15], the hyperparameters were chosen. The mini-batch Stochastic Gradient Descent (SGD) was chosen to learn the network parameters. The batch size was set to 32 (due to limited GPU memory) and the momentum to 0.9. The weights were initialized from a network pretrained on kinetics dataset [16]. Two stream networks were trained where the first stream was RGB and the other for optical flow.

The training procedure starts with temporal learning training, followed by complementary learning to enrich the representation. For training TRN-RGB (TRN for RGB frames), the learning rate was set to 0.001 which decreases every 4000 iterations by $\frac{1}{10}$ with a maximum number of iterations of 11000. On the other hand, on TRN-Optical (TRN for optical flow frames), the learning rate was set to 0.005 which was decreased by $\frac{1}{10}$ at 3000, 7500, and 11500 iterations with a maximum number of iterations of 18000. For T-SOPN-RGB (T-SOPN for RGB frames) and T-SOPN-Optical (T-SOPN for optical flow frames), the learning rate was chosen as 0.001 and 0.005, respectively. It was then reduced by $\frac{1}{10}$ every 10000 iterations and the maximum number of iterations set to 20000 iterations. For GLN, the learned weights from TRN-RGB and TRN-Optical were set as an initialization for the global branch. Then, the whole network was fine-tuned with a learning rate of 0.001 and 0.005, respectively. The learning rate was decreased by $\frac{1}{10}$ after 2000, 3000, and 4000 iterations with a maximum number of iterations of 5000.

Similarly, FIN's fusion branch was initialized from TRN-RGB and TRN-Optical. Then, the whole network was fine-tuned. The learning rate was set for RGB and optical flow to 0.001 and 0.005, respectively. It was decreased it every 1000 iterations by $\frac{1}{10}$. The maximum number of iterations was set to 3000. It is worth noting that during training both GLN and FIN, TRN was incorporated. In other words, a CRF layer was added to GLN and FIN networks for temporal modelling.

Our models were trained using a single GPU nvidia TITAN Xp graphics card. optical flow was extracted using TVL1 algorithm [129].

5.4.3 Ablation Study

The effect of every choice including the network architecture, the number temporal partitions, temporal learning technique, and complementary learning technique was investigated.

The Choice of Number of Partitions and Network Architecture

The choice of the number of partitions in temporal learning is crucial as it governs how dense the long term temporal information is. In the experiments, the number of partitions is varied and the results of comparison between the recognition accuracy of TRN-RGB on UCF101 dataset using two network architectures: Inception V1 [51] and Inception V3 [130] and summarized the results in Table 5.1. The number of partitions N is varied from N = 1..9 — N is odd. It is observed that increasing the number of partitions has a positive influence on the performance. However, N = 7 partitions has the highest performance of 91.09 %. which is 2 % higher recognition accuracy than

Table 5.2: Exploration of Pooling Techniques of InceptionV1 on UCF101 dataset (RGB). TSOP-V1 refers to TSOP based on Eq. 5.10, TSOP-V2 refers to simplified TSOP and AVG refers to average pooling.

Pooling	Recognition Accuracy %
TSOP-V1	90.5 %
TSOP-V2	91.43 %
AVG	89.77 %

Inception V3. Thus, Inception V1 is chosen with TRN (7 partitions).

Table 5.1: Exploration of Changing the Number of Partitions on UCF101 Dataset.

N	Inception V1	Inception V3
1	88.5 %	83%
3	90.8 %	84.6%
5	90.54 %	87.96 %
7	91.09%	88.07 %
9	90.8 %	86.5 %

The Choice of Temporal Pooling Technique

One of the crucial parameters of T-SOPN-RGB based network is the choice of temporal pooling. First, TSOP pooling based on Eq. 5.10 (TSOP-V1), Simplified TSOP (TSOP-V2) and average pooling are compared on RGB videos. The experiments on UCF101 dataset using Inception V1 are conducted after dividing the video sequence into N =7 partitions and the results tabulated in Table 5.2. It is the results tabulated from the table that TSOP-V2 achieved the highest recognition accuracy 91.43 % which is around 1 % higher than TSOP-V1. These results highlight TSOP-V2 as a good candidate to capture temporal information.

Table 5.3: Comparison Between T-SOPN and TRN of InceptionV1 on UCF101 Dataset for both RGB and Optical Flow.

Modality	TRN	T-SOPN
RGB	91.1 %	91.43%
Optical flow	92.57 %	84.45 %

Temporal Learning Technique Vs. The Modality

T-SOPN (Based on TSOP-V2) and TRN for both RGB and optical flow on UCF101 dataset are evaluated. The results are summarized in Table 5.3. From the table, T-SOPN-RGB has slightly higher recognition accuracy than TRN-RGB achieving 91.43 %. However, TRN-Optical achieved recognition accuracy of about 8 % over T-SOPN-Optical. These results imply the effectiveness of TRN and performance consistency against different modalities.

Complementary Learning via GLN

To check the improvement introduced by GLN along with TRN-RGB with N = 7 partitions, several candidates for squeeze and excitation block are evaluated: including AVG pooling only [125], combined MAX and AVG pooling with the same fully connected layers (Shared), and combined MAX and AVG pooling with two separate fully connected layers (Separate). The experimental results are summarized in Table 5.4 on UCF101 dataset. The results show the effectiveness of applying MAX-AVG pooling instead of solely applying AVG. Additionally, MAX-AVG (Separate) has the highest accuracy of 92 % which is higher than AVG pooling only by around 0.7 %.

Furthermore, two architectures shown in Figure 5.6 are investigated. The results are summarized in Table 5.5 on UCF101 dataset. From the table, it is observed that architecture 1 outperforms architecture 2, implying the effectiveness of having two separate branches.

Table 5.4: Exploration of different Pooling techniques for the local branch on UCF101 Dataset (RGB). "MAX-AVG (Shared)" refers to sharing the same fully connected layers between MAX and AVG pooling. "MAX-AVG (Separate)" refers to having two different fully connected layers for MAX and AVG pooling.

Squeeze Pooling	Recognition Accuracy
AVG	91.35%
MAX-AVG (Shared)	91.62 %
MAX-AVG (Separate)	92.0 %

Table 5.5: Exploration of Different Architectures of GLNs on UCF101 dataset (RGB).

Network	Recognition Accuracy
Architecture 1	92.0%
Architecture 2	91.93 %

Complementary Learning Via Fuse-Inception Network (FIN)

The learned representation from backbone 1 is summed with the representation from backbone 2. However, it is crucial to locate the fusion point in the network. Here, three possible locations for the case of Inception V1 are evaluated. Table 5.6 summarizes the results of the experiments conducted on UCF101 dataset. It is observed that after the inception module (4e), the highest performance compared to the rest. Therefore, the decision was made to fuse the output of branch 1 after inception module (4e) shown in Figure 5.7 with the output of branch 2 after inception module (4e).

Table 5.6: Exploration of Fusion point at different points on UCF101 Dataset (RGB).

Fusion Point	Recognition Accuracy
After Inception module (5b)	91.25%
After Inception module (5a)	91.77%
After Inception module (4e)	92.12%

5.4.4 Comparison Against State-of-The-Art

After analyzing the effect of the complementary and temporal learning techniques, we compare our action recognition accuracy against state-of-the-art techniques. The experiments were conducted on two publicly available datasets including UCF101 and HMDB-51 datasets. The comparison is summarized in Table 5.7. Our best result outperforms other methods by approximately 4 % for HMDB-51 and 0.3 % for UCF101 datasets. The superior performance demonstrates the ability of the proposed framework to capture long term temporal information and importance of complementary learning. We notice that fusing RGB and optical streams (Two streams) improves the results. We can also notice from the table that complementary learning techniques (GLN and FIN) improves the performance of TRN, as the network trained on TRN and complementary learning technique (GLN/FIN) achieves higher performance than solely trained using TRN. Furthermore, it is observed from the results that the performance is improved by fusing FIN and GLN. Additionally, the results show that T-SOPN has information complementary to the representation learned by TRN. The proposed method achieves competitive results against 3D Deep ConvNets based methods as illustrated in Table

5.8. It outperforms all 3D Deep ConvNets based methods except I3D which requires 64 frames from each video in training, a computationally intensive process. Experimental results show that the performance of I3D is highly sensitive to the number of the training frames. When the number of frames is reduced, the performance of I3D drops. For example, when the number of frames is set to 16, the performance of I3D is reduced by around 4 %. On the other hand, TRN-RGB achieved nearly the same recognition accuracy for 7 frames only as shown in Table 5.1. An example of *Run* action video and the top four classified actions by the proposed framework from HMDB-51 dataset is shown in Figure 5.8. From the figure, the second, third and forth highest classified actions are either similar to *Run* (*walk*) or have *Run* as a sub-action *dibbling*.

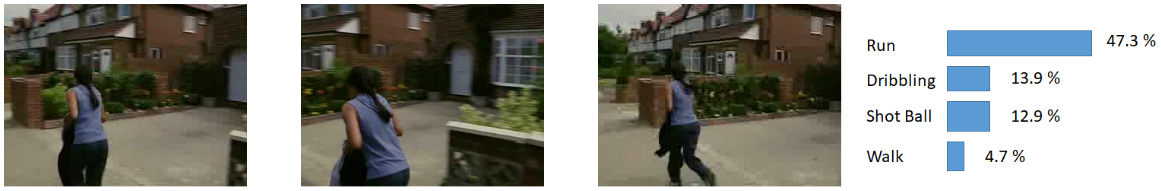


Figure 5.8: An example of ”Run” action video and the top action classes classified by the proposed framework from HMDB-51 dataset.

Table 5.7: Accuracy performance comparison of the proposed method against with state-of-the-art methods based on 2D CNNs.

Technique	UCF101	HMDB-51
Two stream [37]	88.0%	59.4%
IDT [131]	86.4%	61.7%
Dynamic Image Networks [46]	89.1%	65.2%
TDD+IDT [41]	91.5%	65.9%
Two stream Fusion + IDT [38]	93.5%	69.2%
Two stream (TSN) [15]	94.0%	68.5%
Conv Fusion [38]	92.5%	65.4%
Two stream ST-ResNet [39]	93.4%	66.4%
Two stream Spatio Temporal Multiplier [40]	94.2%	68.9%
Deep Temporal Encoding [43]	95.6%	71.1%
Four stream Optical Guided [45]	96.0%	74.2%
Two stream (TRN) Ours	95.34%	72.59%
Two stream (GLN) Ours	95.42%	73.8%
Two stream (FIN) Ours	95.63%	71.8%
Two stream (FIN) + Two stream (GLN) Ours	95.69%	75.14 %
Two stream (FIN) + Two stream (GLN) + T-SOPN Ours	**96.3**%	**76.07**%

Table 5.8: Accuracy performance comparison of proposed method against with state-of-the-art methods based on 3D CNNs.

Technique	UCF101	HMDB-51
C3D[49]	82.3%	56.8%
LTC-RGB-Optical flow [50]	91.7%	64.8%
T3D [59]	93.2%	63.5%
Res3D [52]	85.8%	61.7%
Resnet34 3D [52]	87.7%	59.1%
ResNeXt-101 3D [52]	90.7%	63.8%
Multi Fiber Network [61]	96%	74.6%
I3D-RGB-Optical flow-reported [16]	98.0%	80.7%
I3D-RGB-reported [16]	95.1%	74.3%
I3D-RGB*	91.09%	-
Ours	**96.3 %**	**76.07 %**

5.5 Conclusion

In this chapter, a video based action recognition framework is presented that improved the long term information and harnesses the representation. Temporal learning using Temporal Relational Network (TRN) and Temporal Second Order Pooling based Network (T-SOPN) is proposed. Moreover, complementary learning is introduced via Global-Local Network (GLN) and Fuse-Inception-Net (FIN). The effectiveness of each part is demonstrated through an ablation study. Additionally, the results are compared to other state-of-the-art methods and demonstrate that the proposed technique outperforms all 2D Deep ConvNet based action recognition models.

Chapter 6

Conclusion and Future Work

6.1 Summary

In this dissertation, the problem of human action recognition is investigated. Guided by the goal of improving action recognition performance, three frameworks are proposed. The presented frameworks share the goal of improving the action recognition performance embodying, to some extent, the concept of fusing different sets or modalities. The first two frameworks are based on the fusion of features representing heterogeneous sets or modalities. Moreover, in the second framework, the fusion of deep and hand-crafted features is exploited, giving the representation better resilience against error. The second framework is a transitional framework which combines the merits of deep learning and fusion. However, it is not an end-to-end trained framework which limits the performance. In the third framework, in addition to exploring fusion for enhancing the performance, deep learning is explored specifically for RGB videos which fits the large scale datasets.

First, a human action recognition framework is introduced based on the proposed Multimodal Hybrid Centroid Canonical Correlation Analysis (MHCCCA) which is introduced for two or more sets or modalities. Additionally, a new skeleton representation is proposed to represent the key poses of a subject performing an action. As proof of performance, evaluation is conducted on four publicly available datasets which comprise different modalities, specifically: RGB, depth, skeleton and accelerometer data. The proposed framework showed an improved performance over single handcrafted features, and other fusion methods including Canonical Correlation Analysis (CCA) and Multiset Canonical Correlation Analysis (MCCA). The visualization of the learned features indicates the ability to handle smeared classes in an effective manner.

Second, another human action recognition framework is presented which explores the idea of fusing heterogeneous features, such as handcrafted and deep learning features from different modalities including RGB, depth, and skeleton. For RGB videos, optical flow CNN is adopted as a feature representation capturing the temporal information. Also, HP-DMM-CNN, based on Deep ConvNets is employed as a depth feature representation. At the heart of the proposed framework are the proposed fusion techniques: Multiset Globality Locality Preserving Canonical Correlation Analysis (MGLPCCA) two or more sets or modalities. To showcase the effectiveness of each component of the proposed framework, five datasets are used for testing. The experiments show the robustness of the proposed deep features over their counterparts. Moreover, the proposed fusion techniques achieved better performance over other similar methods. The visualization of the learned features shows the discriminative power of MGLPCCA over CCA.

Third, deep learning based action recognition is explored for large scale datasets. The proposed framework is based on improving the long term temporal information and improve the action representation through temporal and complementary learning. For temporal learning, two possible techniques are introduced, Temporal Relational Network (TRN) and Temporal Second Order Pooling based Network (T-SOPN). The former combines Conditional Random Field (CRF) and 2D Deep ConvNets in order to model different snippets from the video. The latter learns the complementary learning based on pooling feature representation over time using TSOP pooling. Moreover, complementary learning is introduced by introducing Global-Local Network (GLN) and Fuse-Inception Network (FIN). The former learns the complementary information of the global and local representations. The latter encourages the network to learn the complementary information through the learning procedure and loss function. In general, complementary learning improves the final representation through fusion. The effectiveness of each part through an ablation study is demonstrated. The obtained results showed the boosting introduced by the proposed technique to be of high calibre, compared to state-of-the-art methods on large scale datasets.

6.2 Potential Future Work

This dissertation presents three frameworks for human action recognition. The proposed frameworks have room for future improvements.

The first two frameworks, which are based on the fusion of different modalities or sets, can be generalized to other applications. For example, it can be used in fusing features describing the body motions, visual features representing the facial features, and audio features capturing the speech characteristics. Moreover, they can be used for different action recognition problems. For instance, they can be used in cross-view action recognition. In cross-view action recognition, only one view is present at testing and called the target view. On the other hand, two views are present at training. In such case, the algorithm learns to map the representation from source view to target view. Both of the first two frameworks can be adapted for such application.

The first two frameworks can be further improved by introducing neural network based version. Following the recent trend of transformation to a gradient based form of any traditional machine learning technique, two neural network layers can be introduced including MGLPCCA and MHCCCA. In this case, the whole network can be further improved by end-to-end training. Also, these can also leverage the idea of fusing information from different layers within the same Deep ConvNets.

For the third framework, several possible future directions are possible:

1. In this thesis, the temporal learning is applied to 2D Deep ConvNet which can be further improved by applying it to 3D Deep ConvNets. In this case, the temporal long term modelling for the 3D ConvNets is improved by capturing more local and global temporal information regarding the action itself.

2. Moreover, Hidden Conditional Random Field (HCRF) can be adopted instead of Conditional Random Field (CRF) to further enhance the interaction modelling among the sub-actions.

3. Another important future work is to study LSTMs combined with the proposed models which would likely improves the long term temporal modelling especially for action prediction. As attention models introduced by LSTMs showed great performance in other computer vision tasks [132] and have the potential to be applied in temporal modelling in action recognition.

4. CCA can be used as a fusion layer in a ConvNet architecture fusing different possible modalities including RGB, optical flow, and human body poses.

5. The proposed technique can be applied to other similar applications including action anticipation which is the ability of predicting the upcoming action as the proposed method has a better ability of long term temporal modelling.

6. It can be applied as action proposal network which is a core part of action detection, where the main role of the network is to represent the amount of action in a set of frames.

7. Another important future direction is to test the time cost for the proposed techniques and its ability to execute in real-time, a key consideration for deploying any human action recognition system in potential industrial products.

References

[1] D. Aranki, G. Kurillo, P. Yan, D. M. Liebovitz, and R. Bajcsy, "Continuous, real-time, tele-monitoring of patients with chronic heart-failure: lessons learned from a pilot study," in *Proceedings of the 9th International Conference on Body Area Networks*, (London, Great Britain), pp. 135–141, 2014.

[2] V. Bloom, D. Makris, and V. Argyriou, "G3d: A gaming action dataset and real time action recognition evaluation framework," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, (Rhode Island, United States), pp. 7–12, 2012.

[3] Y. Nam, S. Rho, and J. H. Park, "Intelligent video surveillance system: 3-tier context-aware surveillance system with metadata," *Multimedia Tools and Applications*, vol. 57, no. 2, pp. 315–334, 2012.

[4] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 3, pp. 257–267, 2001.

[5] R. D. Green and L. Guan, "Quantifying and recognizing human movement patterns from monocular video images-part i: a new framework for modeling human motion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 2, pp. 179–190, 2004.

[6] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings of the IEEE International Conference on Computer Vision*, (Beijing, China), pp. 1395–1402, 2005.

[7] I. Laptev and T. Lindeberg, "Local descriptors for spatio-temporal recognition," in *International Workshop on Spatial Coherence for Visual Motion Analysis*, (Prague, Czech Republic), pp. 91–103, 2004.

[8] X. Yang, 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*, (Nara, Japan), pp. 1057–1060, 2012.

[9] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in *Proceedings of the European Conference on Computer Vision*, (Florence, Italy), pp. 872–885, 2012.

[10] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naïve-bayes-nearest-neighbor," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, (Rhode Island, United States), pp. 14–19, 2012.

[11] M. Gowayyed, M. Torki, M. Hussein, and M. El-Saban, "Histogram of oriented displacements (hod): describing trajectories of human joints for action recognition," in *Proceedings of the International Joint Conference on Artificial Intelligence*, (Beijing, China), pp. 2466–2472, 2013.

[12] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 1, pp. 51–61, 2015.

[13] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis, an overview with application to learning methods,” *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[14] M. Hasan, “On multi-set canonical correlation analysis,” in *Proceedings of the International Joint Conference on Neural Networks*, (Montreal, Canada), pp. 1128–1133, 2009.

[15] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in *Proceedings of the European Conference on Computer Vision*, (Amsterdam, Netherlands), pp. 20–36, Springer, 2016.

[16] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, United States), pp. 6299–6308, 2017.

[17] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, “Action recognition from depth maps using deep convolutional neural networks,” *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 498–509, 2016.

[18] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, “Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, (Phoenix, United States), pp. 3697–3703, 2016.

[19] I. Laptev, “On space-time interest points,” *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[20] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the Fourth Alvey Vision Conference*, (Manchester, United Kindom), pp. 147–151, 1988.

[21] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, (Beijing, China), pp. 65–72, 2005.

[22] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proceedings of the European Conference on Computer Vision*, (Marseille, France), pp. 650–663, 2008.

[23] P. R. Beaudet, "Rotationally invariant image operators," in *Proceedings of the 4th International Joint Conference on Pattern Recognition*, (Tblisi, USSR), p. 579–583, 1978.

[24] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Proceedings of the IEEE International Conference on Computer Vision*, (Kyoto, Japan), pp. 104–111, 2009.

[25] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features," in *Proceedings of the IEEE International Conference on Computer Vision workshops*, (Kyoto, Japan), pp. 514–521, 2009.

[26] J. Shi and C. Tomasi, "Good features to track," tech. rep., Cornell University, 1993.

[27] J. Sun, Y. Mu, S. Yan, and L.-F. Cheong, "Activity recognition using dense long-duration trajectories," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, (Singapore), pp. 322–327, 2010.

[28] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (San Francisco, United States), pp. 3360–3367, 2010.

[29] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Alaska, United States), pp. 1–8, 2008.

[30] P. Scovanner, 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*, (Augsburg, Germany), pp. 357–360, 2007.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, (San Diego, United States), pp. 886–893, 2005.

[32] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, "Action recognition by dense trajectories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Colorado Springs, United States), pp. 3169–3176, 2011.

[33] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[34] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification,” in *Proceedings of European Conference on Computer Vision*, (Crete, Greece), pp. 143–156, Springer, 2010.

[35] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (San Francisco, United States), pp. 3304–3311, 2010.

[36] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Columbus, United States), pp. 1725–1732, 2014.

[37] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Proceedings of the Advances in Neural Information Processing Systems*, (Montreal, Canada), pp. 568–576, 2014.

[38] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas, United States), pp. 1933–1941, 2016.

[39] C. Feichtenhofer, A. Pinz, and R. Wildes, “Spatiotemporal residual networks for video action recognition,” in *Proceedings of the Advances in Neural Information Processing Systems*, (Barcelona, Spain), pp. 3468–3476, 2016.

[40] C. Feichtenhofer, A. Pinz, and R. P. Wildes, “Spatiotemporal multiplier networks for video action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, United States), pp. 4768–4777, 2017.

[41] L. Wang, Y. Qiao, and X. Tang, “Action recognition with trajectory-pooled deep-convolutional descriptors,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Boston, United States), pp. 4305–4314, 2015.

[42] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, “Real-time action recognition with enhanced motion vector cnns,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas, United States), pp. 2718–2726, 2016.

[43] A. Diba, V. Sharma, and L. Van Gool, “Deep temporal linear encoding networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, United States), pp. 2329–2338, 2017.

[44] S. Kong and C. Fowlkes, “Low-rank bilinear pooling for fine-grained classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, United States), pp. 365–374, 2017.

[45] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, “Optical flow guided feature: a fast and robust motion representation for video action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, United States), pp. 1390–1399, 2018.

[46] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, “Dynamic image networks for action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas, United States), pp. 3034–3042, 2016.

[47] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, “Temporal relational reasoning in videos,” in *Proceedings of the European Conference on Computer Vision*, (Munich, Germany), pp. 803–818, 2018.

[48] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Boston, United States), pp. 2625–2634, 2015.

[49] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, (Santiago, Chile), pp. 4489–4497, 2015.

[50] G. Varol, I. Laptev, and C. Schmid, “Long-term temporal convolutions for action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1510–1517, 2018.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Boston, United States), pp. 1–9, 2015.

[52] K. Hara, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, United States), pp. 6546–6555, 2018.

[53] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas, United States), pp. 770–778, 2016.

[54] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016.

[55] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, United States), pp. 1492–1500, 2017.

[56] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, United States), pp. 4700–4708, 2017.

[57] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, United States), pp. 7794–7803, 2018.

[58] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proceedings of the European Conference on Computer Vision*, (Munich, Germany), pp. 399–417, 2018.

[59] A. Diba, M. Fayyaz, V. Sharma, A. Hossein Karami, M. Mahdi Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3d convnets using temporal transition layer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (Salt Lake City, United States), pp. 1117–1121, 2018.

[60] A. Diba, M. Fayyaz, V. Sharma, M. Mahdi Arzani, R. Yousefzadeh, J. Gall, and L. Van Gool, "Spatio-temporal channel correlation networks for action classification," in *Proceedings of the European Conference on Computer Vision*, (Munich, Germany), pp. 284–299, 2018.

[61] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Multi-fiber networks for video recognition," in *Proceedings of the European Conference on Computer Vision*, (Munich, Germany), pp. 352–367, 2018.

[62] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," in *Proceedings of the Iberoamerican Congress on Pattern Recognition*, (Buenos Aires, Argentina), pp. 252–259, 2012.

[63] C. Liang, L. Qi, Y. He, and L. Guan, "3d human action recognition using a single depth feature and locality-constrained affine subspace coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2920–2932, 2017.

[64] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, (Waikoloa, United States), pp. 1092–1099, 2015.

[65] O. Oreifej 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*, (Columbus, United States), pp. 716–723, 2013.

[66] Q. D. Tran and N. Q. Ly, "Sparse spatio-temporal representation of joint shape-motion cues for human action recognition in depth sequences," in *Proceedings of the RIVF International Conference on Computing and Communication Technologies-Research, Innovation, and Vision for Future*, (Ha Noi, Vietnam), pp. 253–258, 2013.

[67] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, United States), pp. 595–604, 2017.

[68] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, (Rhode Island, United States), pp. 20–27, IEEE, 2012.

[69] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," vol. 25, pp. 24–38, Elsevier, 2014.

[70] A. A. Chaaraoui, J. R. Padilla-Lopez, P. Climent-Perez, and Florez-Revuelta, "Evolutionary joint selection to improve human action recognition with rgb-d devices," *Expert Systems with Applications*, pp. 786–794, 2014.

[71] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Proceedings of the International Joint Conference on Artificial Intelligence*, (Beijing, China), pp. 1351–1357, 2013.

[72] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *Proceedings of the European Conference on Computer Vision*, (Amsterdam, Netherlands), pp. 816–833, 2016.

[73] Y. Du, 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*, (Boston, United States), pp. 1110–1118, 2015.

[74] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proceedings of the ACM on Multimedia Conference*, (Amsterdam, Netherlands), pp. 102–106, 2016.

[75] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, 2017.

[76] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, (New Orleans, United States), pp. 7444–7452, 2018.

[77] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (Columbus, United States), pp. 486–491, 2013.

[78] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Rhode Island, United States), pp. 1290–1297, 2012.

[79] A. Shahroudy, G. Wang, and T.-T. Ng, "Multi-modal feature fusion for action recognition in rgb-d sequences," in *Proceedings of the International Symposium on Communications, Control and Signal Processing*, (Athens, Greece), pp. 1–4, 2014.

[80] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *Proceedings of the IEEE Workshop on Applications of Computer Vision*, (Tampa, United States), pp. 53–60, 2013.

[81] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2123–2129, 2015.

[82] H. Boström, S. F. Andler, M. Brohede, R. Johansson, A. Karlsson, J. van Laere, L. Niklasson, M. Nilsson, A. Persson, and T. Ziemke, *On the Definition of Information Fusion as a Field of Research.* 2007. QC 20180122.

[83] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[84] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Transactions on Systems,Man, and Cybernetics Part B: Cybernetics*, vol. 40, no. 6, pp. 1438–1446, 2010.

[85] M. Caramia and P. Dell'Olmo, *Multi-objective management in freight logistics: Increasing capacity, service level and safety with optimization algorithms.* Springer Science & Business Media, 2008.

[86] J. Bezdek and R. Hathaway, "Convergence of alternating optimization," *Journal Neural, Parallel and Scientific Computations*, vol. 2, no. 3, 2003.

[87] G.Johansson, "Visual motion perception," *Scientific American*, pp. 76–88, 1975.

[88] N. E. D. Elmadany, Y. He, and L. Guan, "Human gesture recognition via bag of angles for 3d virtual city planning in cave environment," in *Proceedings of the International Workshop on Multimedia Signal Processing (MMSP)*, (Montreal, Canada), pp. 1–5, IEEE, 2016.

[89] S. Nowozin and J. Shotton, "Action points: A representation for low-latency online human action recognition," *Microsoft Research Cambridge, Technical Report MSR-TR-2012-68*, 2012.

[90] N. E. D. El Madany, Y. He, and L. Guan, "Human action recognition using temporal hierarchical pyramid of depth motion map and keca," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, (Xiamen, China), pp. 1–6, 2015.

[91] J. Davis, "Recognizing movement using motion histograms," *MIT Technical Report 47*, 1999.

[92] M. Elhoushi, J. Georgy, A. Noureldin, and M. Korenberg, "Motion mode recognition for indoor pedestrian navigation using portable devices," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 1, pp. 208–221, 2016.

[93] S. J. Preece, J. Y. Goulermas, L. P. Kenney, and D. Howard, "A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 871–879, 2008.

[94] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, (San Francisco, United States), pp. 9–14, 2010.

[95] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proceedings of the IEEE International Conference on Image Processing*, (Québec city, Canada), pp. 168–172, 2015.

[96] C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, (Shanghai, China), pp. 2712–2716, 2016.

[97] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[98] C. C. Chang and C. J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.

[99] R. Vemulapalli, 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*, (Columbus, United States), pp. 588–595, 2014.

[100] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807–811, 2016.

[101] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Video emotion recognition with transferred deep feature encodings," in *Proceedings of the ACM on International Conference on Multimedia Retrieval*, pp. 15–22, 2016.

[102] O. Seddati, S. Dupont, S. Mahmoudi, and M. Parian, "Towards good practices for image retrieval based on cnn features," in *Proceedings of the IEEE International Conference on Computer Vision*, (Venice, Italy), pp. 1246–1255, 2017.

[103] Y. Fu, Z. Li, J. Yuan, Y. Wu, and T. S. Huang, "Locality versus globality: Query-driven localized linear models for facial image computing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 12, pp. 1741–1752, 2008.

[104] Y. Yuan and Q. Sun, “Multiset canonical correlations using globality preserving projections with applications to feature extraction and recognition,” *IEEE Transactions on Neural Network and Learning Systems*, vol. 25, no. 6, 2014.

[105] X. He and P. Niyogi, “Locality preserving projections,” in *Proceedings of the Advances in Neural Information Processing Systems*, (Vancouver, Canada), pp. 153–160, 2004.

[106] S. Huang, A. Elgammal, J. Lu, and D. Yang, “Cross-speed gait recogntion using speed invariant gait templates and globality-locality preserving projections,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 10, 2015.

[107] T. Sun and S. Chen, “Locality preserving cca with applications to data visualization and pose estimation,” *Journal Image and Vision Computing*, vol. 25, no. 5, 2007.

[108] J. K. Aggarwal and L. Xia, “Human activity recognition from 3d data: A review,” *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[109] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” in *Proceedings of the British Machine Vision Conference*, (Nottingham, United Kingdom), pp. 1–12, 2014.

[110] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” in *Proceedings of the European Conference on Computer Vision*, (Prague, Czech Republic), pp. 25–36, 2004.

[111] G. Gkioxari and J. Malik, “Finding action tubes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Boston, United States), pp. 759–768, 2015.

[112] X. Xing, K. Wang, and Z. Lv, “Fusion of gait and facial features using coupled projections for people identification at a distance,” *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2349–2353, 2015.

[113] D. Huang, S. Yao, Y. Wang, and F. De La Torre, “Sequential max-margin event detectors,” in *Proceedings of the European Conference on Computer Vision*, (Zurich, Switzerland), pp. 410–424, 2014.

[114] S. Gaglio, G. L. Re, and M. Morana, “Human activity recognition process using 3-d posture data,” *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 5, 2014.

[115] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, “Two-person interaction detection using body-pose features and multiple instance learning,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, (Rhode Island, United states), pp. 28–35, IEEE, 2012.

[116] H. Rahmani, D. Q. Huynh, A. Mahmood, and A. Mian, “Discriminative human action classification using locality-constrained linear coding,” *Pattern Recognition Letters*, vol. 72, pp. 62–71, 2016.

[117] N. E. D. Elmadany, Y. He, and L. Guan, “Multiview emotion recognition via multi-set locality preserving canonical correlation analysis,” in *Proceedings of the IEEE International Symposium on Circuits and Systems*, (Montreal, Canada), pp. 590–593, 2016.

[118] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, “Towards good practices for very deep two-stream convnets,” *arXiv preprint arXiv:1507.02159*, 2015.

[119] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, United States), pp. 1325–1334, 2018.

[120] K. Yu and M. Salzmann, "Statistically-motivated second-order pooling," in *Proceedings of the European Conference on Computer Vision*, (Munich, Germany), pp. 600–616, 2018.

[121] S. Hou, X. Liu, and Z. Wang, "Dualnet: Learn complementary features for image recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, (Venice, Italy), pp. 502–510, 2017.

[122] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the International Conference on Machine Learning*, (Williamstown, United States), pp. 282–289, 2001.

[123] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, United States), pp. 3076–3086, 2017.

[124] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, (Santiago, Chile), pp. 1449–1457, 2015.

[125] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, United States), pp. 7132–7141, 2018.

[126] L. Zheng, Y. Zhao, S. Wang, J. Wang, and Q. Tian, "Good practice in cnn feature transfer," *arXiv preprint arXiv:1604.00133*, 2016.

[127] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[128] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, (Barcelona, Spain), pp. 2556–2563, 2011.

[129] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Proceedings of the Joint Pattern Recognition Symposium*, (Berlin, Germany), pp. 214–223, Springer, 2007.

[130] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas, United States), pp. 2818–2826, 2016.

[131] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, (Sydney, Australia), pp. 3551–3558, 2013.

[132] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the International conference on machine learning*, pp. 2048–2057, 2015.