DESIGN AND DEVELOPMENT OF CONVOLUTIONAL NEURAL NETWORK (CNN) ENABLED SMART CAMERA BASED USER ENTRANCE INTENTION DETECTION SYSTEM

LAI YING CHAU

This Report Is Submitted In Partial Fulfilment Of Requirement For The Bachelor Degree of Electronic Engineering (Computer Engineering)

Fakulti Kejuruteraan Elektronik Dan Kejuruteraan Komputer
Universiti Teknikal Malaysia Melaka

June 2017

"Saya akui laporan ini adalah hasil kerja saya sendiri kecuali ringkasan dan petikan yang tiap-tiap satunya telah saya jelaskan sumbernya."

Tandatangan : .....
Nama Penulis : LAI YING CHAU
Tarikh : 2 JUN 2017

"Saya/kami akui bahawa saya telah membaca karya ini pada pandangan saya/kami karya ini adalah memadai dari skop dan kualiti untuk tujuan penganugerahan Ijazah Sarjana Muda Kejuruteraan Elektronik (Komputer)."

Tandatangan : .....
Nama Penyelia : DR. LIM KIM CHUAN
Tarikh : 2 JUN 2017

To my parents, Lai Son Been and Tan Yoke Har,
my supervisors Dr Lim Kim Chuan and Dr Soo Yew Guan.
and my family and friends.

ACKNOWLEDGEMENT

Firstly, I would like to thank my family especially my father, Lai Son Been and my mother, Tan Yoke Har. Without their love and support over the years none of this would have been possible. They have always been there for me and I am thankful for everything they have helped me achieve.

Next, I would like to thank my supervisor Dr Lim Kim Chuan and Co-supervisor Dr Soo Yew Guan for your generous help and guidance over the years which is unmeasurable and without it I would not be where I am today. I thank you so much for the knowledge you have passed on and I will always be grateful for having the opportunity to be supervised and study under you.

I would also like to express my thanks to the deanery of Faculty of Electronic and Computer Engineering and head of Computer Engineering Department for their support ship to the student of higher education, the faculty is irreplaceable and their generosity to the student body is incomparable.

Last, thank to my colleagues and friends for giving such massive passion to me as there were hard times and we overcame it together. Also to all those have significantly contributed directly or indirectly towards the completion of this final year project. I am truly grateful to all as always being nice and cooperative to me.

ABSTRACT

This project aims to design and develop a system of Convolutional Neural Network enabled smart camera to detect user intention of entering a door based on the user's walking behaviour analysis. Convolutional Neural Network used in this project is 3D Convolutional Neural Network (C3D) integrating with Long Short Term Memory(LSTM). A total of 198 videos, various type of user's behaviour for walking by with and without entering the access door are collected and added into the training and testing dataset. The collected video resolution is 720P with 29 frames per second and the test image size is 640×360. The dataset is then distributed into 50% for training, 25% for validation and 25% for testing. The videos features extracted from the C3D is used as input to train the Recurrent Neural Network (RNN) that learns to classify video clips of 16 frames. It is first trained to identify walking human and the behaviour of entering a door, entering an access door by scanning an access card and passing by. After clip prediction, the output of the RNN is being post-processed to assign a single activity label to each video, and determine the temporal boundaries of the activity within the video. The experimental results show that the CNN model achieves 99.77% highest accuracy of prediction score and 100% of prediction accuracy.

ABSTRAK

Projek ini bertujuan untuk mereka dan membina sistem pengesan niat pengguna untuk memasuki pintu dengan menggunakan kamera pintar berdasarkan *Convolutional Neural Network (CNN)* berdasarkan analisis kelakuan berjalan kaki pengguna. CNN yang digunakan dalam projek ini adalah jenis *3D Convolutional Neural Network (C3D)* dengan integrasi *Long Short Term Memory (LSTM)*. Sebanyak 198 video terdiri daripada pelbagai tingkah laku berjalan pengguna dengan memasuki dan tanpa memasuki pintu akses telah dikumpul dan dimasukkan ke dalam set data latihan dan ujian. Resolusi video yang dikumpul adalah 720p dengan 29 bingkai sesaat dan saiz ujian imej adalah 640 × 360. Set data ini kemudiannya diasingkan sebanyak 50% untuk latihan, 25% untuk pengesahan dan 25% untuk ujian. Kemudiannya, ciri-ciri video yang dihasil daripada C3D akan dimasukkan sebagai input *Recurrent Neural Network (RNN)* untuk melatih dan mengklasifikasi 16 bingkai klip video. Pertama sekali, model C3D dan RNN ini dilatih untuk mengenal pasti pejalan kaki dan tingkah laku pengguna ketika memasuki pintu, memasuki pintu akses dengan mengimbas kad akses dan melalui pintu. Selepas proses ramalan klip video, output RNN *post-process* adalah untuk menetapkan label aktiviti dalam setiap video, dan seterusnya menentukan sempadan aktiviti dalam video. Keputusan eksperimen menunjukkan bahawa model CNN ini telah mencapai ketepatan skor ramalan tertinggi sebanyak 99.77% dan ketepatan ramalan sebanyak 100%.

TABLE OF CONTENT

CHAPTER	TITLE	PAGE
	PROJECT TITLE	I
	DECLARATION	III
	APPROVAL	IV
	DEDICATION	V
	ACKNOWLEDGEMT	VI
	ABSTRACT	VII
	ABSTRAK	VIII
	TABLE OF CONTENT	IX
	LIST OF TABLES	XII
	LIST OF FIGURES	XIII
	LIST OF ABBREVIATIONS	XVI
I	INTRODUCTION	1
	1.1 Introduction	1
	1.2 Problem Statement	3
	1.3 Objectives	3
	1.4 Scope	3
	1.5 Chapter Review	4
II	LITERATURE REVIEW	5
	2.1 Human Gait Recognition	5
	2.1.1 Gait Energy Image (GEI)	6

		2.1.2 Conventional CNN based Gait Recognition	7
		2.1.3 Siamese Neural Network	8
		2.1.4 3D Convolutional Neural Network	9
		2.1.5 Long Short Term Memory(LSTM)	11
	2.2	Graphics Processing Unit (GPU)	11
	2.3	Summary	13
III		METHODOLOGY	16
	3.1	Inferencing the recorded user entering behavior video with the default C3D	16
		3.1.1 CNN Architecture Selection	16
		3.1.2 Data Collection for User Walking Behavior Videos	17
	3.2	Training C3D with User Entering and Pass by Behavior Videos	20
		3.2.1 Modified Architecture of 3D-CNN and RNN Pipeline	20
		3.2.2 System State Transition Diagram	21
	3.3	Project Implementation Flowchart	21
		3.3.1 Extract Features from C3D	23
		3.3.2 Create Stateful Dataset	25
		3.3.3 Train	26
		3.3.4 Predict	28
		3.3.5 Post Process	28
	3.4	Running the Trained C3D on NVIDIA JETSON TX1 Embedded Board	28
IV		RESULTS	31

	4.1	Inferencing the recorded user entering behavior video with the default C3D	31
	4.2	Training C3D with User Entering and Pass by Behavior Videos	32
	4.3	Running the Trained C3D on NVIDIA JETSON TX1 Embedded Board	37
V		CONCLUSION AND RECOMMENDATION	39
	5.1	Conclusion	39
	5.2	Recommendation	40
		REFERENCES	41

LIST OF TABLES

NO	TITLE	PAGE
2.1	Summary Table for Literature Review	13
3.1	C3D Architecture	24
4.1	Summary of results of experiments conducted	34

LIST OF FIGURES

NO	TITLE	PAGE
2. 1	Classification of gait recognition system	5
2.2	Siamese Neural Network Model	8
2.3	The framework of Siamese neural network based gait recognition for human identification by Zhang	9
2.4	Extraction of multiple features from contiguous frames	11
2.5	GPU-accelerated computing	12
3.1	Architecture of 3D-CNN and RNN pipeline	17
3.2	Wireless IP Camera (P2P)	17
3.3	Collect dataset at entrance door with authentication access	18
3.4	Examples of videos recorded for dataset	19
3.5	Architecture of the proposed user entrance intention detection pipeline to be running in the GPU accelerated CCTV embedded camera platform	20
3.6	System State Transition Diagram	21
3.7	Project Implementation Flowchart	22
3.8	Illustration of a Stateful Dataset	25
3.9	Consequence of a high learning rate which leads to failure of reaching minimum loss	27
3.10	NVIDIA JETSON TX1 with relay connected to GPIO module	29
3.11	GPIO module on NVIDIA JETSON TX1	29
3.12	Jetson TX1 J21 Header Pinout	29
4.1	Result of top 5 classification with default C3D and dataset. (a) lady is cleaning a sink	31

(b) Result of prediction
4.2 Results of inferencing a user entering behaviour video to C3D. 32
(a) a man is walking pass by the door
(b) Wrong prediction of the video of a man passing by the door.
(c) a lady is walking towards a door
(d) Wrong prediction of the lady walking towards a door.
4.3 Graph of C3D and RNN Training Accuracy 33
4.4 Graph of C3D and RNN Training Loss 33
4.5 Example of classification testing video. 35
(a) a video of a man walking toward a door with facing the door and raising his hand is tested by C3D.
(b) Entering activity is predicted with the score of 94.13%.
4.6 Example of classification testing video 36
(a) a video of a lady walking toward a door with facing the door and raising her hand is tested by C3D.
(b) Entering with access activity is predicted with the score of 99.77%.
4.7 Example of classification testing video 36
(a) a video of a man walking directly straight without go near to the door is tested by C3D.
(b) Pass by activity is predicted with the score of 99.46%.
4.8 Example of classification testing video 37
(a) a video of a group of ladies are walking directly straight without go near to the door is tested by C3D.
(b) Pass by activity is predicted with the score of 67.09%.
4.9 Example of classification testing video 37
(a) a video of a lady with carrying a baby and a child who wants to enter an access door is tested by C3D.
(b) Entering with access activity is predicted with the score of 43.48%.
4.10 LED of relay lights up indicate operation of door 37

LIST OF ABBREVIATIONS

C3D	-	3D Convolutional Neural Network
CCTV	-	Closed-circuit Television
CNN	-	Convolutional Neural Network
CPU	-	Central Processing Unit
GEI	-	Gait Energy Image
GPIO	-	General Purpose Input Output
GPU	-	Graphics Processing Unit
HOG	-	Histogram of Oriented Gradient
IPSA	-	Intelligent Personal Security Assistant
LSTM	-	Long Short Term Memory
MSE	-	Mean squared error
P2P	-	Wireless IP Camera
RNN	-	Recurrent Neural Network
SIFT	-	Scale Invariant Feature Transform

CHAPTER I

INTRODUCTION

1.1 Introduction

Recognizing human actions in real-world environment finds applications in a variety of domains including intelligent video surveillance, customer attributes, and shopping behaviour analysis. User entrance intention detection is mainly to indicate a human intention to go through the entrance via observing the human gait and the corresponding movement trajectory. Since the advancement of electronic sensors technology, contemporary sensor-based automatic entrance door control technologies such as infrared and ultrasonic sensors are widely used in public places to convenient the public society from manually opening and closing actions. [1] The infrared sensor based entrance can be further divided into two categories such as active and passive approaches. The active approach emits infrared signals from the controller and receive the reflected signals to indicate objects near the door. This approach is accurate and able to identify the position and speed of object however with higher cost which impact to be less popular in the society. While the passive approach detects infrared signals radiated by people. In comparison, the passive way of infrared sensor based entrance door is widely used for being effective and low cost. On the other hand, ultrasonic sensor based entrance door emits ultrasonic or radio waves to scan the environment and analyses the reflected signals.

These techniques are all successful in controlling automatic entrance door by detecting object approaching near to the door, however they are not capable to understand the type and the intention of the approaching objects. For instance, a puppy or a passing pedestrian may accidentally trigger the door and cause a false opening action. Frequent false opening action of automatic entrance door would impact is air conditioning energy waste and reduces equipment lifetime.

This calls for the need of a user entrance intention detection system based smart camera integrating with automatic door control system since access doors are commonly equipped with CCTV to record the access to the guarded facilities. With the advancement of computer vision for self-driving car technology and accelerated embedded computing platform, the feasibility of using these matured technologies in creating Intelligent Personal Security Assistance (IPSA), running right inside the CCTV camera to determine user entrance intention is being explored in this project.

Human action recognition in videos attracts increasing research interests in computer vision community due to its potential applications in video surveillance, human computer interaction, and video content analysis. It is the main techniques in order to run an Intelligent Personal Security Assistance inside a CCTV camera to detect user entrance intention. Intelligent Personal Assistants, also known as virtual human assistants are devices or software applications that can assist the user with in daily activities, navigation, information retrieval, and organisation.

The recent advances in video coding, storage and computational resources have boosted research in the field towards new and more efficient solutions for organizing and retrieving video content. A simple pipeline composed of a 3D-CNN that exploits spatial and short temporal correlations followed by a Recurrent Neural Network (RNN) which exploits long temporal correlations in order to classify and temporarily classify the activity in untrimmed videos.

1.2 Problem Statement

Current sensor-based automatic entrance door access could lead to false opening action as it is not capable to understand the type of objects and the intention for accessing the entrance. This false opening action could waste air conditioning energy and reduce equipment lifetime. While conventional automatic door with authentication access cause inconvenience to the user when the user is entering while carrying a handful of things and it is also inefficiency in flashing access card frequently.

1.3 Objectives

The aim of this study is to design and development of Convolutional Neural Network enabled smart camera based user intention detection system for the purpose of creating Intelligent Personal Security Assistance (IPSA). Some objectives are needed to be accomplished in order to achieve the aim of this study.

a) To identify suitable CNN based technique for classifying user intention (body language) of approaching and accessing the access door.
b) To train the identified CNN (with transfer learning) to automatically learn body language (gait and trajectory) from video sequences to classify the intention of user.

1.4 Scope of work

In this project, a CNN based algorithm with human action recognition in videos will be developed for door control system which is based on the confirmation that the detected object is a human and the corresponding gait and movement trajectory indicates that he/she has the intention to access the entrance by implementing the use of convolutional neural network for learning and extracting high-level features from the video captured by a smart camera installed above the entrance door. If the human is approaching the door and has an intention of entering, the door should unlock automatically without going through identity authentication. A sufficient number of

human entering or passing by an entrance content videos dataset is recorded in order to achieve high accuracy and good performance in human gait recognition in videos.

1.5 Chapter Review

Chapter 1 describes the general overview of this project. This chapter presents the introduction, problem statement, objectives, scope of work and review of all chapters included in this thesis.

Chapter 2 discusses the terminology of deep learning techniques being used in this project. The techniques to be discussed includes the 3D-CNN that exploits spatial and short temporal correlations and Recurrent Neural Network (RNN) which exploits long temporal correlations. This chapter also presents the previous and related work to this project.

Chapter 3 presents the methodology to develop algorithms for assigning a label for the input video, recognizing activities in untrimmed videos and find the temporal segments in which the required human gait appears in the video. Chapter 4 delivers the results obtained in this studies and discussion. Chapter 5 gives the conclusion and recommendation of this project.

CHAPTER II

LITERATURE REVIEW

2.1 Human Gait Recognition

Human gait recognition can be classified into three groups namely, motion vision based, wearable sensor based and floor sensor based. The motion vision can be divided into two groups namely; appearance based methods and model based methods. The appearance based method can be also subdivided in two types; state space methods and spatio-temporal methods [2, 3, 4, 5]. Figure 2.1 shows motion vision based system flow chart.

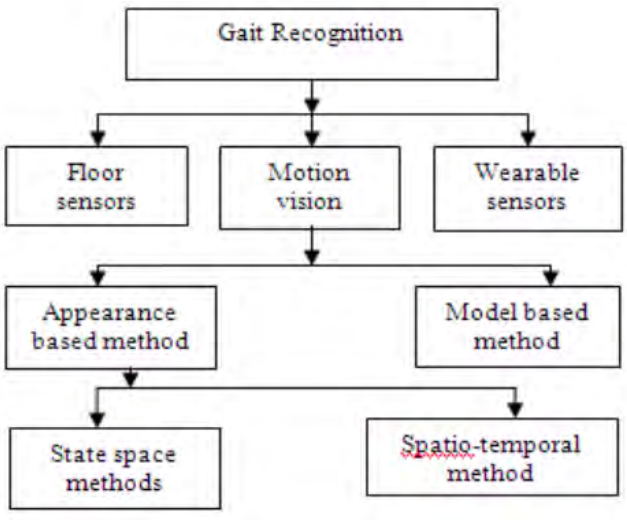

Figure 2.1: Classification of gait recognition system

Data representation plays the main role in affecting the performance of learning algorithm. Thus, average silhouette approaches for a sequence was introduced by Liu and Sarkar [6]

2.1.1 Gait Energy Image (GEI)

The Gait Energy Image (GEI) representation was proposed by Han and Bhanu [7]. Gait Energy Image (GEI) is the sum of images of the walking silhouette divided by the number of images. GEI is a useful representation with superior selective power and strength against segmental errors [8]. It has been widely used due to its simplicity and effectiveness. According to a recent study by Iwama, et al, which evaluated the performance of different gait representation templates using a large gait dataset (containing no covariates) showed that it outstands from other gait representation templates as it saves storage space and computation time for recognition.

Gait Energy Image (GEI) method to apply recursive principle component analysis technique achieved better recognition rate [9, 10, 11, 12]. Gait Energy Image (GEI) representation were used for different application purposes such as removing subject backpack without losing subject original shape and information [9], human identification [10], person recognition [11] and recognition on gender or age groups [13]. Lee et al. applied to their method to normal walk, slow walk, and fast walk for experiment. They have used CASIA C dataset for conducting the test.

Shingh and Biswas approached gait energy image (GEI) method for human identification. They selected normal walk, wearing a coat or jacket or carrying a bag for recognition purposes. They informed that normal walk sequence is obtaining better recognition rate compared to carrying a bag or wearing a jacket or coat. They focused on subject body alignment with bottom and upper part of the body as feature. They also reported that gait recognition rate can be improved by applying GEI method. They selected large CASIA gait database for the experiment [10, 14].

Ju and Bir proposed gait energy image (GEI) method for person recognition individually. They created statistical gait features from actual and artificial gait

templates for the experiment. They selected USF HumanID gait database for gait recognition purposes. They also used others gait database to compare recognition rate with current method with selected gait database. The GEI method is obtained better recognition rate after comparing published gait result [11].

Okumura et al., (2010) described a large-scale gait database that can use widely for vision based gait recognition. They focused on gait energy image method for recognition on gender or age groups. From the experiment, female subjects are achieving better recognition rate compared to male [15] . For the age grouping, it's evaluated according to maturity of walking ability and also physical strength. They have got different fluctuation from different age groups. They also compared with several gait databases to evaluate their method performance [13].

2.1.2 Conventional CNN based Gait Recognition

The hand-crafted Conventional CNN used the traditional image processing method for example: Histogram of Oriented Gradient (HOG) and Scale Invariant Feature Transform (SIFT). Conventional CNN based gait recognition is an attempt to fine-tune the conventional CNN on the gait dataset for gait recognition by achieving below tasks:

a) CNN is capable to learn discriminative features automatically by exploring deep architecture at multiple level of abstracts from raw data, without any domain knowledge.
b) Fine-tuning from a pre-trained model is a good solution to solve the data limitation problem and speed up the convergence of new model.

As in [16], they fix all convolutional layers and fine-tune the fully connected layers. After training, the activations of three fully connected layers (CNN.FC1, CNN.FC2 and CNN.FC3) as the feature representations and implement the KNN method to identify the person in surveillance environment. However, the conventional CNN based method unable to solve the gait recognition effectively as conflict between large categories number and small samples per category occurred.

These hand-crafted systems dominate the object recognition field and are very useful for the object recognition on small scale datasets. However, the drawbacks of hand-crafted system are obvious when it is employed with large scale datasets as it requires millions of templates for template matching. Different from hand-crafted systems, deeply-learned systems learnt the image features from the images automatically. While the Convolution Neural Network (CNN) shows its great performance for ImageNet image classification.

2.1.3 Siamese Neural Network

The Siamese Network was first introduced by Yann LeCun et al. in year 1993 to be applied to face and signature verification task [17]. Siamese network is used to learn a function that maps input patterns into latent space where similarity metric to be small for pairs of the same objects and large for pairs from different objects. This is implemented by feeding the output of two sub identical network that share the same weights and parameters to a cost function that compute the distance measure by using Cosine Similarity function.

Figure 2.2 shows the Siamese neural network model. Thus, Siamese network is the best suited for verification scenarios such as gait recognition where the number of classes is very large and examples of all the classes are not available at the time of training.

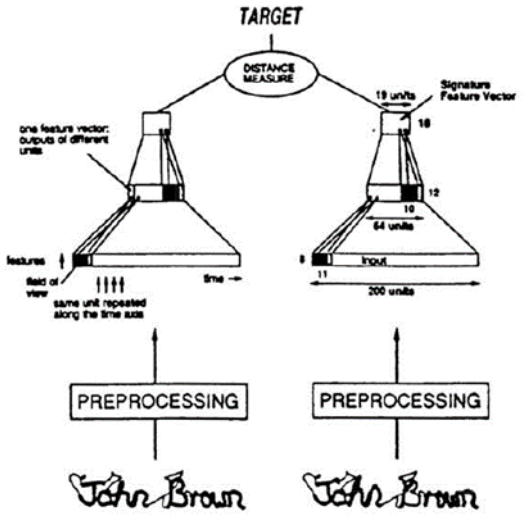

Figure 2.2: Siamese Neural Network Model

Siamese Neural Network designed by Zhang contains two parallel CNN architectures, which consist of two parts, the two convolutional layers and max-pooling layers and three fully connection layers as shown in Figure 2.3 [16]. In the training stage, the two branches of the network will be optimized simultaneously with the weight sharing mechanism. Pairwise images with similar or dissimilar labels separately entrance the two CNNs. Then the output of the CNNs are combined by the contrastive layers to compute the contrastive loss. After that, the back-propagating with contrastive loss is used to fine-tune the model. They use OULP-C1V1-A-Gallery dataset, with 20,000 similar GEI pairs and randomly selected 20,000 dissimilar pairs for training sets while send the query GEI into one of the CNNs during testing stage to compute feedforward network based on the matrix multiplication for one time to extract features.

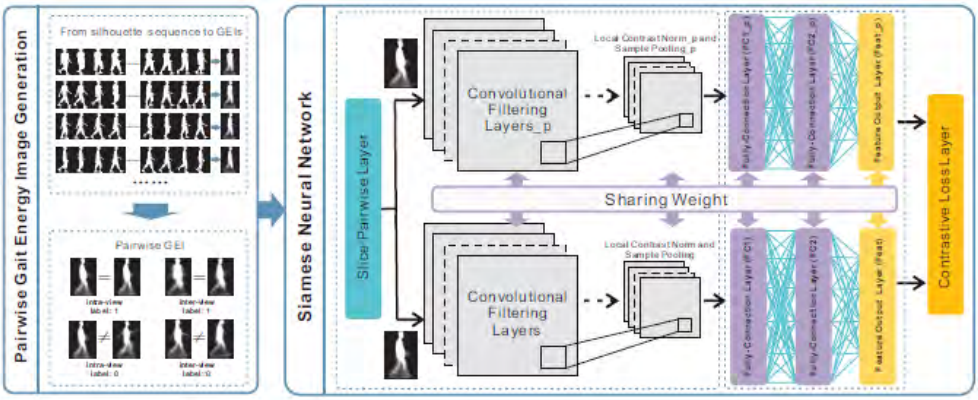

Figure 2.3: The framework of Siamese neural network based gait recognition for human identification by Zhang. *[16]*

2.1.4 3D Convolutional Neural Network

Convolutional neural network (CNN) is a deep model that can employ directly on the raw inputs, thus automating the process of feature construction. CNN models can be categorized into 2D CNN and 3D CNN. 2D CNN model is performed at the convolutional layers to extract features from local neighborhood on feature maps in the previous layer. Then an additive bias is applied and the result is passed through a sigmoid function. Formally, the value of unit at position (x, y) in the jth feature map in the *ith* layer, denoted as $v_i^{xy}{}_j$, is given by

$$v_{i\ j}^{xy} = \tanh\left(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right) \qquad (2.1)$$

Where tanh(.) is the hyperbolic tangent function, b_{ij} is the bias for this feature map, *m* indexes over the set of feature maps in the (*i*-1)th layer connected to the current feature map, w_{ijk}^{pq} is the value at the position (*p,q*) of the kernel connected to the *k*th feature map, and *Pi* and *Qi* are the height and width of the kernel, respectively. In the subsampling layers, the resolution of the feature maps is reduced by pooling over local neighborhood on the feature maps in the previous layer, thereby increasing invariance to distortions on the inputs. A CNN architecture can be constructed by stacking multiple layers of convolution and subsampling in an alternating fashion. The parameters of CNN, such as the bias b_{ij} and the kernel weight w_{ijk}^{pq}, are usually trained using either supervised or unsupervised approaches [18, 19]. While 3D CNN extracts feature from both spatial and temporal dimensions [20] by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames [21]. This is necessary to capture the motion information encoded in multiple contiguous frames when applied to video analysis problems. The 3D convolution is achieved by convolving a 3D kernel to the cube formed by stacking multiple contiguous frames together. By this construction, the feature maps in the convolution layer is connected to multiple contiguous frames in the previous layer, thereby capturing motion information.

$$v_{i\ j}^{xyz} = \tanh\left(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{Ri-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \qquad (2.2)$$

where *Ri* is the size of the 3D kernel along the temporal dimension, w_{ijm}^{pqr} is the (*p, q, r*)th value of the kernel connected to the *m*th feature map in the previous layer.

As the kernel weights are replicated across the whole cube, only one type of features can be extracted from the frame cube. Multiple 3D convolutions can be applied to contiguous frames to extract multiple features. As in Figure 2.4, the sets of connections are color-coded so that the shared weights are in the same color. Note that