ACTION RECOGNITION FOR SOCIAL ROBOTS

by

Rahmetullah VAROL

B.S., Mechatronics Engineering Department, Yildiz Technical University, 2016

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering Department
Boğaziçi University

2020

ACKNOWLEDGEMENTS

I would like to express my deepest graditute to my thesis supervisor Prof. Levent Akın, whose guidance made this dissertation possible. I want to thank the dissertation committe members Assoc. Prof. Hatice Köse and Assist. Prof. Fatma Başak Aydemir for their brilliant suggestions and comments. I would like to thank my parents Ahmet Varol and Tuba Varol and my sisters Fatma Betül Varol, Hatice Kübra Varol and Zeynep Varol whose support was one of primary sources of my motivation. Finally, I would like to thank my beloved Zeynep Karavelioğlu who stood by me through all the ups and downs of my life since the day we met.

ABSTRACT

ACTION RECOGNITION FOR SOCIAL ROBOTS

Being able to understand and recognize actions is a crucial precondition for social integration, which enables the members of a social community to interact with each other and with their environments. During the development of a person's cognitive abilities, the ability to detect and recognize actions improve over the course of a long period. First, we learn to recognize simple and explicit actions that have little ambiguity, such as; waving, eating, or walking. As we progress this ability we learn to recognize subtler actions, such as; smiling, resting, or reading. In more complex cases, these actions may overlap or compose a more integral action, such as riding a bike. In a similar spirit to humans, for a robot to be able to make natural and seamless interactions with people and make plans that are appropriate for the state of the environment it is in, it is imperative that it can understand the actions of the people around it. For this, a robot needs a powerful action recognition module that can on real-world conditions where a variety of distortions are present. Action recognition is the task of observing the sequential progression of these movements and matching some segments of this sequence with previously defined action classes which have been labeled by the action type that defines them. In many cases, the task of action recognition comes together with the task of action detection. Action detection is the task of extracting the segments from a usually long observation that contains some action. In many cases, action recognition and detection occur naturally in humans and we subconsciously recognize a person's actions without much effort. This ability makes social interaction much smoother. However, this is a very challenging problem for computer, since many actions contain complex action segments that might have very similar appearances and temporal processions. Even subtle differences can put an action into an entirely different class.

ÖZET

SOSYAL ROBOTLARDA AKSIYON TANIMA

İnsan aksiyonlarını tanıma ve anlayabilme sosyal integrasyon için önemli bir önşarttır ve sosyal bir topluluğun üyelerinin birbirleri ile ve çevreleri ile etkileşim kurmalarına olanak sağlar. Bir insanın algısal kabiliyetlerinin gelişimi sırasında, insan aksiyonlarını tespit etme ve tanıma kabiliyeti uzun bir zaman içerisinde gelişir. İlk olarak; basit, belirgin ve içerisinde fazla belirsizlik içermeyen aksiyonları tanımayı öğreniriz. Bunlara örnek olarak el sallama, yemek yeme veya yürüme aksiyonları verilebilir. Bu kabiliyetimizi geliştirdikçe, gülümsemek, dinlenmek veya okumak gibi incelikli aksiyonları da algılamaya başlarız. İnsanlara benzer bir şekilde, robotlar için de, etrafındaki insanlarla doğal ve akıcı etkileşim kurabilmek ve içinde bulunduğu duruma göre hareketlerini planlayabilmek, etrafındaki insanların hareketlerini algılayabilme ve anlayabilme kabiliyeti önem arz etmektedir. Bunun için, sosyal bir robot farklı bozucu etkilerin mevcut olduğu gerçek hayat koşullarında çalışabilen güçlü bir aksiyon tanıma modülüne ihtiyaç duymaktadır. Aksiyon tanıma, insan hareketlerinin zamana bağlı olarak gözlenmesi ve bu gözlemlerin önceden tanımlanmış bazı aksiyon sınıfları ile eşleştirilmesidir. Çoğu durumda, aksiyon tanıma ve aksiyon tespit etme görevlerine eş zamanlı olarak ihtiyaç duyulmaktadır. Aksiyon tespit etme, genellikle uzun bir görüntüden içinde tekil aksiyonlar bulunan parçaların tespit edilmesidir. Aksiyon tanıma ve aksiyon tespit etme insanlarda doğal olarak gerçekleşir ve bilinçaltımızda insanların aksiyonlarını tanıma işlemini fazlaca efor sarf etmeden gerçekleştiririz. Bu kabiliyet sosyal etkileşimlerimizi daha akıcı hale getirir. Fakat, bilgisayarlar için bu zorlu bir görevdir, çünkü birçok aksiyon karmaşık aksiyon parçalarının bir araya gelmesi ile oluşur ve aksiyonlar oldukça benzer görünümlere ve zamansal bağlamlara sahip olabilir. Küçük farklılıklar dahi insan hareketlerini farklı bir aksiyon sınıfına yerleştirmek için yeterli olabilir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . ix
LIST OF TABLES . . . xi
LIST OF SYMBOLS . . . xii
LIST OF ACRONYMS/ABBREVIATIONS . . . xiv
1. INTRODUCTION . . . 1
1.1. Action Detection and Recognition . . . 2
1.1.1. Definition of Action in the Literature . . . 2
1.1.2. Definition of Action Used In This Thesis . . . 3
1.1.3. Action Localization . . . 4
1.1.4. Action Recognition . . . 5
1.2. Action Recognition in Social Robotics . . . 6
1.2.1. Social Robots and Their Place in Society . . . 6
1.2.2. Action Recognition in the Context of Social Robotics . . . 8
1.3. Motivation of This Thesis . . . 9
1.4. Aims and Objectives of This Thesis . . . 11
1.5. Overview of Methodology . . . 12
1.5.1. Action Proposals . . . 12
1.5.2. Transfer Learning of Temporal Segment Networks . . . 13
1.6. Contributions . . . 14
2. LITERATURE REVIEW . . . 15
2.1. Research on the Psychological Aspects of Action Recognition . . . 15
2.2. A Taxonomy of Action Recognition Methods . . . 16
2.2.1. Action Taxonomies in the Literature . . . 16
2.2.2. Action Taxonomy Used in This Thesis . . . 18
2.3. Representation Based Action Recognition Techniques . . . 18

2.3.1. Holistic Representation Based Methods . . . 18
2.3.1.1. Space-Time Approaches . . . 18
2.3.1.2. Sequential Approaches . . . 19
2.3.1.3. Non-Hierarchical Single Layer Approaches . . . 19
2.3.2. Local Descriptors Based Action Recognition Techniques . . . 19
2.3.2.1. SIFT Descriptor Based Models . . . 19
2.4. Deep Neural Network Based Action Recognition Techniques . . . 20
2.4.1. Spatiotemporal Networks . . . 20
2.4.2. Multiple Stream Networks . . . 20
2.4.3. Deep Generative Networks . . . 21
2.4.4. Temporal Segment Networks . . . 21
2.5. Datasets . . . 21
2.5.1. Simple Action Datasets . . . 22
2.5.1.1. KTH Human Action Dataset . . . 22
2.5.1.2. Weizmann Human Action Dataset . . . 22
2.5.2. Complex Action Datasets . . . 22
2.5.2.1. MSR Action Dataset . . . 22
2.5.2.2. HMDB-51 Dataset . . . 23
2.5.2.3. JHMDB Dataset . . . 23
2.5.3. Internet Datasets . . . 24
2.5.3.1. UCF YouTube Action Dataset . . . 24
2.5.3.2. ActivityNet Dataset . . . 24
2.5.4. RGB-D Datasets . . . 25
2.5.4.1. CAD-60 and CAD-120 . . . 25
3. ACTION LOCALIZATION via DEEP TEMPORAL ACTION PROPOSALS 26
3.1. Problem Definition . . . 26
3.2. Action Localization in a Real-Time and Online Setting . . . 27
3.2.1. Single Shot Detector Model . . . 27
3.2.2. Real-Time Optical Flow Computation . . . 28
3.2.3. Bounding Box Prediction . . . 29
3.3. Experiments and Results . . . 30

3.3.1. Evaluation Metric . . . 30
3.3.2. Experiments . . . 30
3.3.3. Extracted Snippets . . . 31
3.3.4. Network Performance . . . 31
4. TEMPORAL SEGMENT NETWORKS for ACTION RECOGNITION . . . 35
4.1. Two Stream Convolutional Neural Networks . . . 35
4.2. Temporal Structure Modeling . . . 36
4.3. Temporal Segment Network Architecture . . . 37
4.3.1. Training of TSNs . . . 38
4.3.2. Testing of TSNs . . . 38
5. TRANSFER LEARNING of TSNs via ACTION PROPOSALS . . . 39
5.1. Retraining the Output Layer . . . 39
5.2. Experiments and Results . . . 39
6. CONCLUSION . . . 41
REFERENCES . . . 42

LIST OF FIGURES

Figure 1.1: The relation between the components of a traditional action recognition framework as given in [5]. At first, features are extracted from raw video frames and then these features are used to learn and segment the actions using the underlying statistical patterns in the data. These statistical patterns are then matched with existing actions in a model database. 6

Figure 1.2: According to this hypothesis, simplistic and unrealistic robots evoke better emotional response on humans than robots that imperfectly resemble actual human beings. 7

Figure 1.3: Some examples of the social robots that are currently being used for various purposes. (a) The companion robot BUDDY developed by the Blue Frog Robotics company as an emotional companion robot, (b) the robot Pepper developed by the SoftBank Robotics company as a customer service robot, (c) the robot Sophia developed by the Hanson Robotics company as a technology demonstration robot, (d) the robotic dog AIBO developed by the Sony Corporation as a companion robot, (e) the humanoid robot NAO developed by the SoftBank Robotics company as a research robot and (d) the robot Kismet developed at the Massachusetts Institute of Technology by Dr. Cynthia Breazeal as a research robot. 10

Figure 1.4: Deep Action Proposal network architecture. 13

Figure 2.1: Some examples of the actions used in the HMDB-51 dataset. 23

Figure 2.2: Screenshots of the 11 actions of UCF YouTube Action Dataset. . . . 24

Figure 3.1: The SSD network. The SSD architecture is based on a feed-forward convolutional neural network which provides bounding boxes and scores for object class instances followed by a non-maximum suppression step. 27

Figure 3.2: Two stream action detection network is based on integrating the single shot detector network architecture into a two-stream convolutional neural network architecture. As such, both the appearance and motion content are considered and fused together to obtain highly confident temporal and spatial action proposals. 32

Figure 3.3: Exemplary snippets that were extracted by the method proposed in this papers. Using these snippets instead of uniformly sampled snippets increases the performance of action Temporal Segment Network performance. 33

Figure 3.4: Action localisation results using the mAP (%) metric on UCF101-24 and JHMDB, at IoU thresholds of $\delta = 0.25, 0.5$. 34

Figure 4.1: Two-stream convolutional neural network architecture used for action recognition. The idea behind this network comes from neurological studies where it was discovered that human brain contained two separate pathways for object recognition and motion recognition. 35

Figure 4.2: Temporal segment network acrhitecture. The two-stream network is parallelized for randomly selected snippets from the whole video. This way temporal structured in the motion and appearence content can be recognized. 36

LIST OF TABLES

Table 2.1: The two dimensional taxonomy of action recognition methods given in [5] by Weinland et al. 17

Table 3.1: IoU comparison with state-of-the-art methods on the HMDB-51 dataset. 32

Table 3.2: IoU comparison with state-of-the-art methods on the THUMOS dataset. 32

Table 5.1: Performance comparison of the proposed TSN+AP network with current state-of-the-art networks [98]. 40

LIST OF SYMBOLS

A	Predicted bounding box
b	Additional parameter of a softmax function
B	Ground-truth bounding box
c_t	Ground-truth action class
C	Number of action classes
$d_i(x)$	Intensity difference
E_G	Gradient energy term
E_I	Sum of intensity energy term
E_S	Smoothness energy term
$E(u)$	Weighted sum of energy terms
F	A function representing a convolutional neural network
G	Segmental consensus function
H	Action class prediction function
I_k	A grayscale frame from an action video
K	Number of segments in an action video
L	Loss function
N_g	Total number of temporal annotations
N_s	Number of patches
p	Ground-truth action class probability density function
S_k	A segment of an action video
$t_{s,n}$	Starting time for the nth temporal annotation
$t_{e,n}$	Ending time for the nth temporal annotation
T_k	An action proposal snippet
$T(x)$	Objective function
$\mathbf{u}$	A warping vector
$\mathbf{U}_k$	Dense flow field
U_s	Weighted averaged dense flow field
v	Warping vector parameter

$\mathbf{W}$	Set of convolutional neural networks parameters
$\mathbf{x}$	Location in the reference image
x_n	nth frame in a given video
X	A given action video.
y_i	Ground-truth label regarding class i
Z	Normalization coefficient
α	Coefficient of intensity term
β	Coefficient of gradient term
γ	Coefficient of smoothness term
δ	Size of an action snippet
θ_{ps}	Patch size
θ_{sd}	Downscaling quotient
$\lambda_{i,x}$	
τ_p	Timestamp pair set
ϕ_n	nth elements of the timestamp pair set
Ψ	Variational refinement function

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
AP	Average precision
CNN	Convolutional Neural Network
DAP	Deep Action Proposal
GAN	Generative Adversarial Network
IoU	Intersection over union
KLT	Kanade-Lukas-Tomasi
mAP	Mean average precision
PM-GAN	Paralleled Mix-Generator Generative Adversarial Network
RGB	Red, green and blue
RGB-D	Red, green, blue and depth
ROC	Receiver-operator characteristic
SIFT	Scale Invariant Feature Transform
SSD	Single Shot Detector
TSN	Temporal Segment Network

1. INTRODUCTION

This thesis investigates the problem of automated human motion recognition in the context of social robotics research. Analysis of human motion is one of the most challenging tasks in computer vision research and there is an extensive literature on this topic. As humans, the ability to analyse and recognize human motion is a critical part of our daily communications and we accomplish this task without much effort. However, for a computer that relies only on a visual feedback, this is a very challenging task. There are many reasons for this, such as the complex interactions that humans build with other humans and/or objects and the wide variety of different contexts human motion can occur. Despite the complexity of this task, it is a worthy endeavour due to the great number of applications for which understanding of human motion is crucial.

Such applications include; monitoring of people in public places for out of ordinary behaviour, detection of pedestrians that may endanger the safety of traffic, monitoring the behaviour of drivers for signs of sleep deprivation etc. In each of these applications, an understanding of human motion is required. In addition to these applications which are mostly centered around monitoring of people for a specific purpose, there are also applications that require the analysis of videos that already exist. Today, millions of hours of video are produced on the internet every single day. For many platforms, automated understanding the human activities that is present in a given video would be a very valuable tool so as to develop tools that can understand the contents of a given video. One interesting application is the summarization of sports events which require detection of certain activities that commonly occur in sporting events. In this study, the application of action recognition in the context of social robots is investigated. This is a challenging task since social robots commonly require interaction with humans in complex social settings.

1.1. Action Detection and Recognition

1.1.1. Definition of Action in the Literature

Before we can define what action recognition is in a meaningful way, we first need to define what an action is. Even though defining what an action seems to be straightforward task, there are many different definitions of action in the literature.

- In [1], Turaga et al. define action as a short pattern of motion executed by a single person. In contrast, an activity is defined as a complex sequnece of actions performed by several people. This definition makes the distinction between an action and an activity based on the number of people and the complex interactions they build in the given context.
- In [2], Poppe defined an action as a whole-body movement that might or might not be cyclic. Additionally, an activity is defined as a series of subsequents actions which gives an interpretation of the movement that is being performed.
- In [3], Chaaraoui et al. an action is defined as primitive movements that can last up to several minutes. An activity is defined as a larger scale event in which the environment and interactions come into the play. According to their definition activities require tracking and classifications of actions in a particular order.
- In [4], Chaquet et al. defined an action as a sequence of primitive actions that serves a simple purpose, such as jumping or walking. An activity however is defined as a series of actions that have a spatial and temporal relation.
- In [5], Weinland et al. defined an action as a sequence of primitive movements realized by a human during the performance of a task. This definition has a similar spirit to the definition of Chaquet et al. in the sense that the center of focus is on the intention behind the movements.
- In [6], Wang et al. relates action to the transformations they impose on the environment. In this context, an action is related to its interaction with the objects around such as kicking a ball. This definition puts the emphasis on the spatial and temporal relations between objects and agents.

1.1.2. Definition of Action Used In This Thesis

In this study we use the definition of action as presented in [7], where Moeslund and Granum define the action hiearchy as action/motor primitives, actions and activities. These terms are described as follows;

Action primitive: The first assumption is that the human motion can be decomposed into atomic bodily movements called action primitives. There are many studies in the literature that support this thesis one example of which is [8] by Schaal et al. where human motion is modeled using an alphabet of simple movements called action primitives. These lay the foundations on which all the more complex structures are modeled.

Action: Actions are defined as a collection of several action primitives that have some spatial and temporal relations among them. These collections can be described at the limb level so as to make it easier to identify and localize singular action instances.

Activity: Activities are defined as larger scale events that are composed of several actions and interactions with the environment. The activities are largely dependent on the context of the environment, objects, or interacting humans. The distinction between action and activity is complex. However, activities are more complex and the activity recognition methods usually are accompanied by object recognition or context recognition methods.

The primary reason for using this definition is the hierarchical approach used when constructing each term. Each term is described as the collection of a set of simpler parts. Such an approach is useful in the context of social robotics since it allows actions to be decomposed into their components.

1.1.3. Action Localization

Action localization, or as sometimes called action segmentation or action detection, comes in two varities. First is the task of spatial action localization and second is temporal action localization. Temporal action localization, is the task of extracting the segments from a usually long observation that contain some action. This task is usually achieved by marking the start and end points of an action in a given video sequence. One commonly used method of marking such points is to use the space-time interest points proposed by Laptev and Lendeberg at [9]. Spatial action localization, is the task of finding the region in a given frame which contains an action. This task usually accompanies its temporal counterpart by detecting the regions of actions in subsequent frames from which action tubes are generated [10].

There is no hard evidence in neuroscientific studies that indicate action recognition and action detection are performed seperately in the human mind. However, from a computational perspective, it is easier to find actions in a given long and untrimmed video if the temporal and spatial limits are known. As such action localization is usually used as a preprocessing step in order to increase the speed or accuracy of the action recognition methods.

Action localization is a challenging task due to the ambiguous and relative nature of this task. When recognizing actions we normally don't put hard boundaries around them and usually it is not clear which body movements should be included in which action sequence. Further, actions that take place in uncontrolled settings usuaaly have parts that are occluded, can be observed from unusual viewpoints, can happen slow or fast at different instances. These ambiguities naturally make the computation and assessment of action detection hard.

1.1.4. Action Recognition

Action recognition is the task of observing the sequential progression of these movements and matching some segments of this sequence with previously defined action classes which have been labeled by the action type that define them. In many cases, the task of action recognition comes together with the task of action detection.

In many cases, action recognition and detection occur naturally in humans and we subconsiously recognize a person's actions without much effort. This ability makes social interaction much smoother. However, this is a very challenging problem for computer, since many actions contain complex action segments that might have very similar appearances and temporal processions. Even subtle differences can put an action into an entirely different class.

Traditionally, action recognition is composed of a several components each of which serves as a preparation step to the next one. In [5], these components are listed as feature extraction, action segmentation and action learning and classification and the relation between these components are given as shown in Figure 1.1. These components are described as follows,

Feature extraction: This component is the preprocessing step to detect the posture and motion cues from a given video in order to extract discriminative features that can be used for identifying the related action content. There are many different features used in the literature from very simple silhoutte images to complex body models.

Segmentation: Action segmentation is the process of cutting the video into segments that contain singular action instances and can be named as a single entity.

Learning and classification: Learning action classes is the step of identifying the statistical patterns in the extracted features and matching the segmented actions

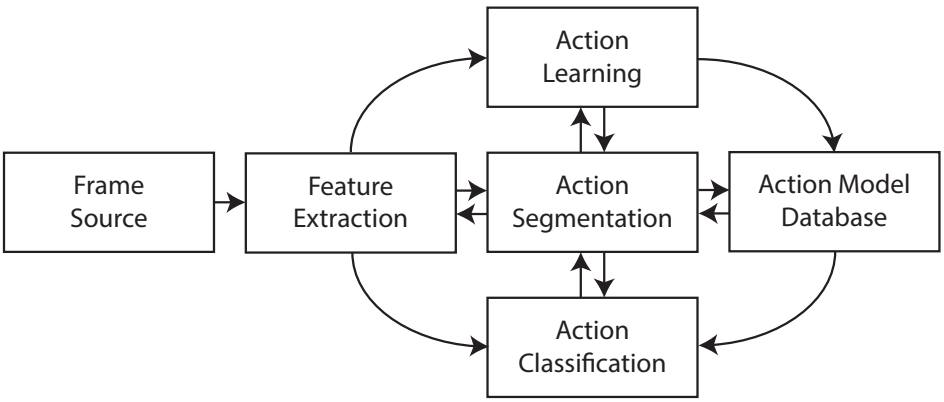

Figure 1.1. The relation between the components of a traditional action recognition framework as given in [5]. At first, features are extracted from raw video frames and then these features are used to learn and segment the actions using the underlying statistical patterns in the data. These statistical patterns are then matched with existing actions in a model database.

to existing classes of actions.

1.2. Action Recognition in Social Robotics

1.2.1. Social Robots and Their Place in Society

A social robot is a physical agent usually in the form of a human or an animal that is designed specifically to interact with humans in such a way that the interaction occurs in a natural way and doesn't irritate the human. In [11], Dautenhahn and Billard proposed the following definition for a social robot;

> Social robots are embodied agents that are part ofa heterogeneous group: a society of robots or hu-mans. They are able to recognize each other and engage in social interactions, they possess histories(perceive and interpret the world in terms of their own experience), and they explicitly communicate with and learn from each other.

Such social requirements require robots to possess various capabilities and use different techniques such as social learning, imitation, emotion and gesture recognition. In accordance with the capabilities of a given robot it may be placed in different

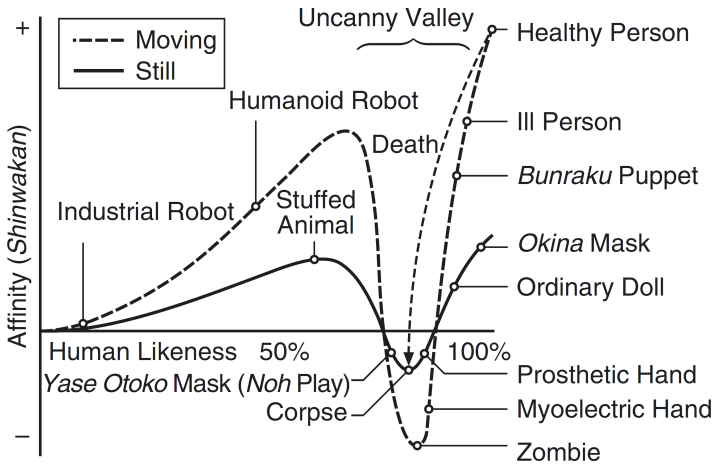

Figure 1.2. According to this hypothesis, simplistic and unrealistic robots evoke better emotional response on humans than robots that imperfectly resemble actual human beings.

classes in a social context. In [12], Breazeal defines four classes of robots according to their social capabilities; socially evocative robots, social interface robots, socially receptive robots and sociable robots. Socially evocative robots are designed to engage in interactive activities with humans and evoke their tendency to empathize with the antromorphic similarities the robot possesses. These robots are mostly used for entartainment or social engagement purposes. Social interface robots on the other hand are intended for more practical uses such as guidance robots or server robots. These robots are socially interactive at the interface level such as speech, gesture and mobility. For these robots, clear communication is more iportant and the psychological ascepts of this communication is not very critical. Socially receptive robotos are situated such that they benefit from the interactions they engage with humans. These robots typically have built-in models of human behavior so as to understand and learn from these interactions. Sociable robots on the other hand have internal social aims that are pre-defined and they actively engage with humans in order to satisfy these aims. In [13], Fong et al. define three more classes as socialy situated robots, socially embedded robots and socially intelligent robots.

Social robots are used for many purposes such as; elderly care, ambient living assistant for disabled people, social service agent at restaurants and other places. For

a social robot to behave appropriately in these settings, it needs to have some form of understanding of the actions the people around it performs. Some examples of the social robots that are currently being used for various purposes. (a) The companion robot BUDDY developed by the Blue Frog Robotics company. BUDDY is specifically designed to express emotions through a screen on its face and develop an emotional connection with its owner. In (b), the Pepper robot developed by the SoftBank Robotics company. Pepper is designed for personal communication and customer service purposes and is generally used by various companies that want to incorporate robots into their customer services. In (c), the robot Sophia is shown. Sophia is a humanoid robot developed by the Hanson Robotics company. Sophia is designed to have a realistic human-like apperance (d) the Robotic dog AIBO (e) the humanoid robot NAO (d) the robot Kismet. All these robots have various appearances from very simplistic and cartoonish robots like the BUDDY robot or to very realistic humanoid robots such as the Sophia robot. An interesting hypothesis about the human likeness in accordance with the visual appearance of the robot which is called the Uncanny Valley theory. This hypothesis was first identified by Mori in 1970 [14]. According to this hypothesis, simplistic and unrealistic robots evoke better emotional response on humans than robots that imperfectly resemble actual human beings. The uncanny valley graph is shown in Figure 1.2.

1.2.2. Action Recognition in the Context of Social Robotics

Action recognition is an indispensable part of social robotics research since the ability of the robot to interpret the events in its vicinity largely depends on its ability of figuring out the activities that are currently taking place. It is probable that social robotics research will be one of the most popular application areas in the following years. However, currently there is limited research on the specifics of action recognition in the context of social robotics research. Some examples are; [15] where Addwiteey et al. investigated the use of different action representations for gesture recognition as part of a social robot that is used inside a conversational context and [16] where Coppola et al. studied the development of a action recognition module that is used for

continous action recognition from a RGB-D (red, green, blue and depth) video stream. While promising results are obtained, there is still much ground to cover in this topic to achieve results that are satisfactory enough for a socially capable robot.

Action recognition is a critical part of a social robot and as such it needs to be precise and fast. One of the challenges of action recognition for social robots in the need for real-time action recognition capabilities.

1.3. Motivation of This Thesis

Robots become more and more integrated into our daily lives with each passing day. They are routinely used in industrial settings such as manufacturing plants [17], search and rescue settings [18], agricultural settings [19], medical settings [20] and in the military [21]. One of the key challenges of robotics research and of the main obstacles in front of large scale human-robot integration is the inadequate social capabilities of current robotic systems. Generally speaking, most of the current robotic systems are operated in isolated and heavily controlled environments. As the capabilities of robots increase they become more and more integrated into the society. However, for a robot to be able to truly function in a society it needs to develop certain social skills. This is where social robotics research come into play.

One of the most important capabilities a robot needs to have is the ability to recognize the actions of an observed human. This ability will allow the robot to plan and act accordingly. Furthermore, it will give the robot the ability to predict future actions of a human and allow it to prepare for upcoming events. The main motivation in this thesis is to develop an action recognition module that is appropriate to be used inside a social mobile robot. Such an ability will allow social robots be more suitable to be used in socially intense settings such as nursing homes, rehabilitation centers or public service centers. Use of social robots in these settings will provide a more interactive setting for the people who need frequent mental stimulation and emotional care. This is an ever increasing need, due to the aging population.

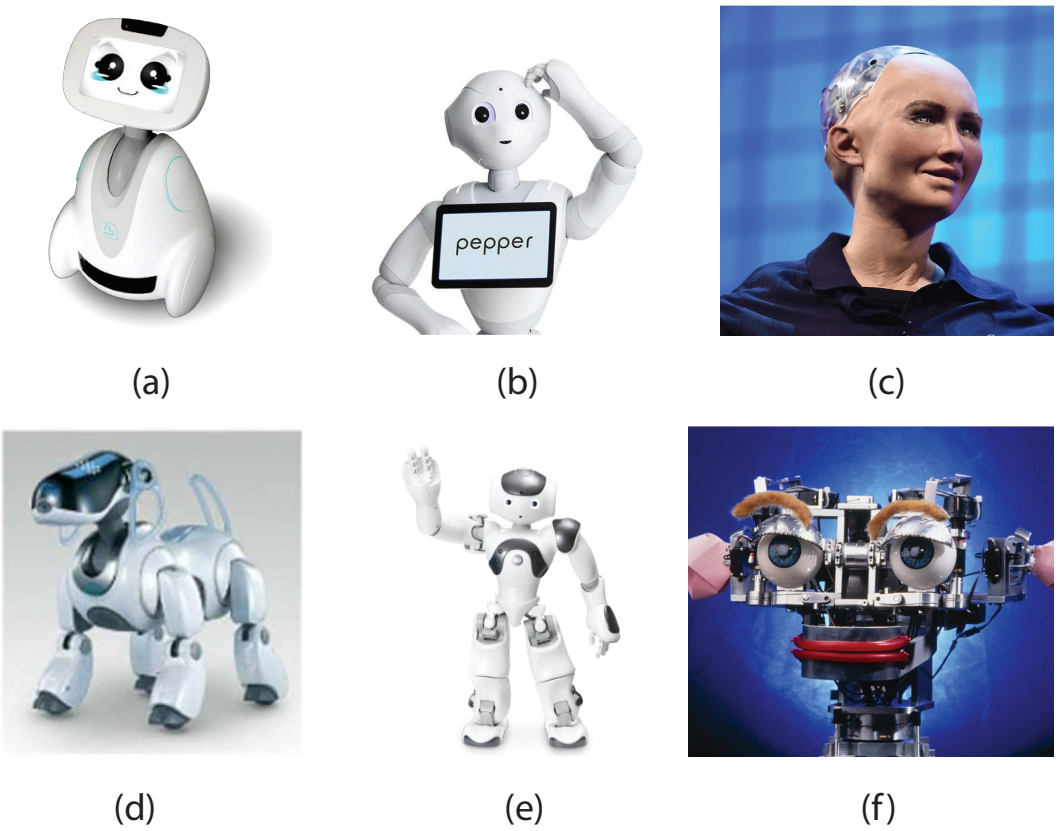

Figure 1.3. Some examples of the social robots that are currently being used for various purposes. (a) The companion robot BUDDY developed by the Blue Frog Robotics company as an emotional companion robot, (b) the robot Pepper developed by the SoftBank Robotics company as a customer service robot, (c) the robot Sophia developed by the Hanson Robotics company as a technology demonstration robot, (d) the robotic dog AIBO developed by the Sony Corporation as a companion robot, (e) the humanoid robot NAO developed by the SoftBank Robotics company as a research robot and (d) the robot Kismet developed at the Massachusetts Institute of Technology by Dr. Cynthia Breazeal as a research robot.

1.4. Aims and Objectives of This Thesis

Being able to understand and recognize actions is a crucial precondition for social integration, which enables the members of a social community to interact with each other and with their environments. During the development of a person's cognitive abilities, the ability of detecting and recognizing actions improve over the course of a long period. First, we learn to recognize simple and explicit actions that have little amibiguity, such as; waving, eating or walking. As we progress this ability we learn to recognize subtler actions, such as; smiling, resting or reading. In more complex cases, these actions may overlap or compose a more integral action, such as riding a bike. In a similar spirit to humans, for a robot to be able to make natural and seamless interactions with people and make plans that are appropriate for the state of the environment it is in, it is imperative that it can understand actions of the people around it. For this, a robot needs a powerful action recognition module that can operate on real-world conditions where a variety of distortions are present.

The aim of this thesis is to develop a multimodal action recognition module for a social robotic assistant that is geared towards recognizing actions that are commonly encountered in a household-like environment. The main motivation behind this study is to increase social capabilities of social assistant robots in order to develop their decision making capabilities in the context of social environments. Regarding the scope of the action classes that will be considering during the training and experimentation process, we chose a subset of the action classes used in ActivityNet [22]. This subset consists mostly of actions that commonly occur inside a household-like environment. The resulting action recognition module will then be implemented on a real social robotic assistant and experiments will be conducted in a household-like environment.

1.5. Overview of Methodology

In this section the framework used in the development on the action recognition module will be introduced. The idea behind the proposed framework is to develop a network architecture that is suitable for modeling the long range temporal relationships in videos. One existing approach to this problem is the use of two-stream convolutional neural networks. Two-stream convolutional neural networks decouple the spatial and temporal information in videos by training two parallel convolutional neural networks with different representations of the video and fusing the output of these parallel networks in a final layer. The idea behind the two-stream neural network architecture is related to the two-stream hypothesis, proposed by Goodale and Milner in [23]. According to this hypothesis, human visual cortex contains two pathways: the ventral stream which performs object recognition and the dorsal stream which recognises motion. In a similar manner, two-stream network contains two streams one of which is trained for recognizing appearance based patterns and the other trained for recognizing motion based patterns.

1.5.1. Action Proposals

Action proposals are a novel and efficient method of obtaining temporal action proposals that extract temporal segments which are likely to contain human actions [24–26]. Similar to grouping techniques for retrieving object proposals, they create a hierarchy of fragments by hierarchical clustering based on semantic visual similarity of contiguous frames. The main disadvantages of this approach are its strong dependence on an unsupervised grouping method that diminishes its repeatability. Recently, a number of papers have been published that extend on this topic [27–29]. One of the prominent examples of these networks is the Deep Action Proposals (DAPs) proposed by Escorcia et al. shown if Figure 1.4, due to its high-performance and real-time implementation capability. This method uses a supervised method that learns to generate segments on a video and predict their action likelihood and leverages long-short term memory cells to learn an appropriate encoding of the video sequence as a set of

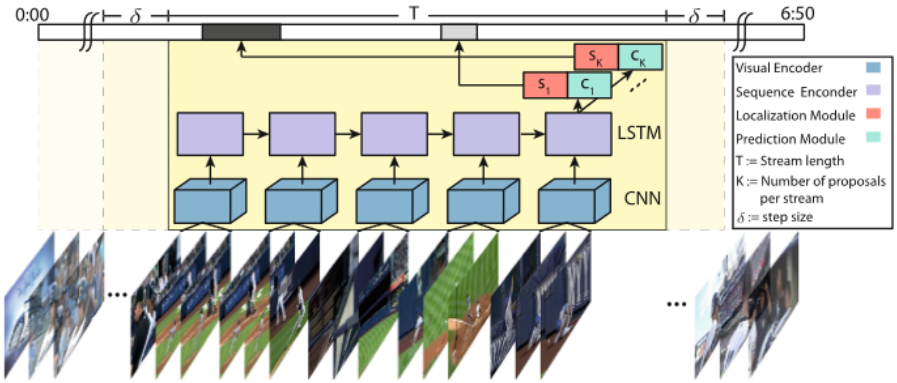

Figure 1.4. Deep Action Proposal network architecture.

discriminative states. The output of this network is a list of short temporal segments that are likely to contain an action. We aim to use the output of this network in the sampling of temporal snippets that are used to train temporal segment networks.

1.5.2. Transfer Learning of Temporal Segment Networks

The main novelty in this thesis is to improve on the sampling strategy of the Temporal Segment Network (TSN) architecture. Current strategy is to divide the vidoe into K segments of equal length and from each of them to ranodmly select a short snippet which will be fed into the network. It is proposed that with a sufficiently high K the selected snippets will be enough represent the content of the whole video. However, for most complex actions this sparse sampling strategy might not be the best method. Because, for most actions particularly the motion content is not uniformly distributed. Some parts of the video weights more than others. For this reason, it is useful to have a prior idea about the parts of the video that has a higher probability of containing motion content.

1.6. Contributions

In this study the main contribution will be to implement an cation recognition module that can be practically used as a sub-module in a social robotic assistant. Since the resulting action recognition module is intended to be used on real-world, conditions being able to operate in real-time and being robust to camera motion and other distortions are crucial criteria it must satisfy. Real-time operating capability will be achieved by strategically sampling potential segments of a continuous video stream.

Recent development in detecting potential frames in which an action might be present will be used for this purpose. Particularly, integration of these networks with recognition modules that have the ability to model temporal relations might yield a powerful methodology that can be used for real-world applications. Another ability we aim for our action recognition module is the ability to operate in the absence of visual feedback. As such, we will try to develop methods that will allow us to recognize patterns of actions from other information sources (e.g. sound or depth).

2. LITERATURE REVIEW

Action recognition has a wide range of applications and has been studied from many perspectives throughout the last couple of decades. The psychological aspects of human motion perception as been studied first. However, in recent years due to the increase in the computational power of our computer and the availability of larger datasets, the computational aspect of action recognition is also being investigated.

The first step in building a good action recognition model is to select an effective feature representation strategy. This is a recurring topic in computer vision research and one of the most widely discussed topics in action recognition research. In recent years, many different action representations methods have been proposed, including local and global features based on temporal and spatial changes trajectory features based on key point tracking [30,31], motion changes based on depth information [32–34], and action features based on human pose changes [33, 35].

2.1. Research on the Psychological Aspects of Action Recognition

Motion perception ability of humans has been studied as early as 1970s starting with the landmark paper in which Johansson showed that humans retain the ability of recognizing actions when the effect of form was removed [36]. In his experiments, he showed that humans can recognize actions from the movement of light sources attached to different joints on the human body. His experiments inspired some of the first techniques in action recognition and similar experiments were performed over the course of several decades. His experiments also led to the discussion on whether the three dimensional (3D) information played an important part in recognizing an action.

The psychological aspects of motion perception and action recognition have been studied widely ever since. Some notable examples are; [37] where Cutting and Kozlowski studied the recognition capabilities in humans of the walking action using only

simple light sources mounted on joints and [38] where Dittrich discussed the implications of Johansson's experiments from the perspective of computer vision, perceptual models and mental representations. However, large scale computational investigation has not been feasible due to limited computational resources and lack of large scale databases suitable for this purpose.

2.2. A Taxonomy of Action Recognition Methods

2.2.1. Action Taxonomies in the Literature

- In [1], Turaga et al. classified both action and activity recognition methods into three different classes. Action recognition methods are divided into non-parametric, volumetric and parametric approaches. Non-parametric approaches are template based where features are extracted from each frame in the video and they are subsequently matched to a stored templated for each action class. Examples of these techniques are; dimensionality reduction, template matching and 3D objects etc. Volumetric approaches consider videos as a 3D matrix of two spatial and one temporal dimension. Features are then extracted in this 3D domain. In these approaches in general extended variants of two dimensional (2D) features are used. Examples of these techniques are; space-time filtering, constellation of parts, sub-volume matching. Parametric approaches on the other hand are model based. These techniques try to capture the temporal dynamics of actions using mathematical models and parameters of these models are then estimated using training data. Examples of these techniques are hidden markov models and linear dynamic systems.
- In [5], Weinland et al. used a two dimensional classification matrix based on how the actions were represented spatially and temporally. In the temporal domain they used the classes; grammar based, template based and temporal statistics based methods. In the spatial domain they used the classes; body model based, image model based and spatial statistics based methods. Thus, nine different classes of actions were devised examples of which are given in Table 5.1.

Table 2.1. The two dimensional taxonomy of action recognition methods given in [5] by Weinland et al.

	Grammars	Templates	Temporal Statistics
Body Models	Body Grammars e.g. Ramanan [40], Green [41], Lv [42], Wang [43], Guerra-Filho [44]	Body Templates e.g. Gavrila [45], Yacoob [46], Rao [47]	Bag of Postures e.g. Ikizler [48]
Image Models	Image Grammar e.g. Elgammal [49], Ogale [50], Turaga [51], Weinland [52], Lv [53], Shi [54], Natarajan [55],	Image Template e.g. Bobick [56], Weinland [57], Laptev [58], Fathi [59], Souvenir [60], Farhadi [61]	Bag of Keyframes e.g. Efros [62], Weinland [63], Schindler [64]
Spatial Statistics	Space Bag of Trajectories, e.g. Messing [65]	Feature Template e.g. Laptev [66], Ke [67]	Bag of Events e.g. Schuldt [68], Boiman [69], Dollar [70], Niebles [71], Niebles [72], Klaser [73], Laptev [9]

- In [39], Herath et al. divided action recognition methods into feature based and learning based methods. This change came after the recent success of deep neural network architectures that don't need any handcrafted features to be extracted. According to this taxonomy, representation based methodologies were divided into holistic representations and local representations. Furthermore, deep network based methodologies were divided into spatio-temporal networks, multiple stream networks, generative models and temporal coherency networks.

  Local representations based methods try to extract local features in individual frames.

2.2.2. Action Taxonomy Used in This Thesis

In this thesis I will use the taxonomy that was proposed by Herath et al. in [39]. The reason behind this is that this taxonomy puts emphasis on more recent methods that depend on deep neural network based architectures.

2.3. Representation Based Action Recognition Techniques

In this section, I will give an overview of the methods that are based on hand-crafted features extracted from action videos.

2.3.1. Holistic Representation Based Methods

In the holistic representations, the strategy is to try to extract a global representations of the whole body structure, shape and movements. These approaches are more likely to preserve the spatial and temporal relationships that are present in the action sequence relative to local descriptors.

2.3.1.1. Space-Time Approaches. Space-time approaches treat time as a regular dimension and extract features from the resulting volumetric three-dimensional data. Some of the work done in this domain use the entire 3D volume as its feature space. These approaches commonly suffer from background noise. Based on [74], [75] proposed to combine motion history images and appearance information in which foreground image (obtained through background subtraction) and histogram of oriented gradients were used as appearance based features. Some other studies use trajectories that are obtained through tracking of joint positions [76], [77]. These studies are mainly based on upon the work by Johansson in [78]. Messing et al. [76] used a Kanade-Lukas-Tomasi (KLT) tracker [79] for tracking Harris3D interest points in order to obtain feature trajectories. Then, a generative mixture model is used to learn a velocity-history language and classify video sequences.

2.3.1.2. Sequential Approaches. Sequential approaches try to capture the temporal relationships in observations. Some of these approaches represent human actions as a sequence of sample set of observations [80], [81]. Some other studies learn a state model for each action and represent them as sets of hidden states [82]. Standard hidden Markov models are commonly used for this purpose.

2.3.1.3. Non-Hierarchical Single Layer Approaches. Non-hierarchical single layer approaches can be used for extracting sub-activities to be used in hierarchical action recognition models. Statistical approaches use statistical models such as dynamic Bayesian networks [83] or conditional random fields [84] to recognize action patterns. Description based approaches extract spatio-temporal descriptions from the video and try to model the sequential and concurrent relationships among them. To this end, Bayesian belief networks [85] and Petri-nets [86] have been used previously.

2.3.2. Local Descriptors Based Action Recognition Techniques

2.3.2.1. SIFT Descriptor Based Models. The scale invariant feature transform (SIFT) methodology is used to generate commonly used descriptors in many computer vision tasks such as object recognition [87] of scene recognition [88]. They have many desirable properties such as scale, transform or rotation independency. Both 2D and 3D SIFT transforms have been used for action modeling and recognition [89, 90]. The two-dimensional variant of the SIFT features are generally used to model the static shape content such as appearance and the 3D variant is used to dynamic content such as motion. There have been interesting demonstrations of the use of SIFT feature for the generation of temporal invariant descriptors [91].

2.4. Deep Neural Network Based Action Recognition Techniques

2.4.1. Spatiotemporal Networks

Spatiotemporal neural networks function similar to convolutional neural networks except that the convolution operation is modified to take advantage of the temporal information. An example of such networks are 3D convolutional neural networks which were developed by Ji et al. in [92].

2.4.2. Multiple Stream Networks

Multiple stream networks are inspired from the two-streams hypothesis of the human visual perception system. This hypothesis was first proposed by Goodale and Milner in their seminal paper [23]. According to this hypothesis, human visual cortex contains two pathways: the ventral stream which performs object recognition and the dorsal stream which recognises motion. In a similar manner, two-stream network contains two streams one of which is trained for recognizing appearance based patterns and the other trained for recognizing motion based patterns. Some variants are also designed such that one stream is performs spatial recognition and one stream performs temporal recognition. Usually the distinction is very subtle. The spatial or dorsal stream is designed to recognize pattern in still frames by identifying the appearance based cues in the frames. Temporal or ventral stream, on the other hand, tries to capture the temporal relations between videos using temporal basde methods such as dense optical flow. In general noth stream are implement as convolutional neural networks. The two-stream hypothesis has been heavily critised and recently it is generally agreed that the visual perception system of humans involve more complex interactions between various perceptive subsystems. A recent critism of this hypothesis can be found at [93]. However, action recognition methods based on the two-stream hypothesis are still popular in the computer vision and machine learning community. Some examples are the two-strema convolutional neural networks developed by Simonyan and Zisserman [94] and by Feichtenhofer et al. [95].

2.4.3. Deep Generative Networks

A recent trend in action recognition research is to utilize generative adversarials networks (GANs) in the task of recognizing human actions. The main objective in using GANs is to try to learn from the temporal relations in action videos in an unsupervised manner. GANs consists of a discriminative and a generative part. The discriminative part is trained using two different types of input one of which comes from a high-dimensional data source and the other is computationally generated random noise. It's objcetive is to discriminate between generated and teal samples. The generative part uses the output of a discriminator and in turn generated better samples. For the fask of action recognitions, GANs are generally used for learning feature representations of actions in videos. Examples of such networks are the DiscrimNet by Ahsan et al. [96] and the Paralleled Mix-Generator Generative Adversarial Network (PM-GAN) by Wang et al. [97].

2.4.4. Temporal Segment Networks

A promising new study by Wang et al. involves a novel architecture called the Temporal Segment Network [98]. This architecture is designed to capture the temporal structures inherent to action videos by randomly sampling short snippets throughout the video. Wang et al. managed to surpass the performance of the state-of-the-art networks in the untrimmed video classification task of ActivityNet challenge.

2.5. Datasets

Until recently there were no large scale datasets that are appropriate for action recognition. Starting with the KTH action dataset [68], a significant advance in the understanding of action recognition methods have been observed. More sophisticated datasets, such as UCF101 [99], Sports-1M [100], HOLLYWOOD2 [101], ActivityNet [22], and others [102], [103] in turn paved the way for more sophisticated algorithms. With the popularization of the use of RGB-D cameras in computer vision domain,

RGB-D datasets are being used more commonly for action recognition research [104], [105], [106]. Using such datasets presents opportunities for developing multimodal algorithms that combine different information sources so as to obtain a robust action recognition system.

2.5.1. Simple Action Datasets

2.5.1.1. KTH Human Action Dataset. The KTH dataset [68] was created at the Royal Institute of Technology, Sweden in 2004. This dataset is the first action dataset that was adopted widely and used as a benchmark. It contains six different action classes. They are performed by 25 different subjects at 4 different scenarios. In tolta there are 600 action videos. Action classes are listed as; walking, jogging, running, boxing, hand waving and hand clapping. Different scenarios are listed as outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4). The camera with which the action were recorded and the background is static. This dataset is one of the simplest action datasets and the highest reported accuracy in this dataset is 97.89% by Moussa et al. at [107].

2.5.1.2. Weizmann Human Action Dataset. The Weizmann Human Action Dataset was created at the Weizmann Instute of Science, Israel in 2005. Similar to the KTH dataset, it is a simple action dataset with 10 actions performed by 9 different actors with a static background. The action classes are bending, jumping jack, jumping, jump in place, running, side jumping, skipping, walking, one-hand and two-hand waving. This dataset is considered as a good benchmark and some studies reported 100% accuracy [108, 109].

2.5.2. Complex Action Datasets

2.5.2.1. MSR Action Dataset. MSR action dataset [110] was created at Microsoft Research in 2009 for the evaluation of the robustness of action recognition methods. In this dataset, the backgrounds are dynamic and cluttered with both outdoor and indoor

Figure 2.1. Some examples of the actions used in the HMDB-51 dataset.

samples. The dataset contains 16 video sequences with 3 types of actions performed by 10 people. The action classes are hand clapping, hand waving and boxing. While being relative small, this is one of the first datasets where actions were recorded in real-life like scenarios.

2.5.2.2. HMDB-51 Dataset. The HMDB-51 dataset was created by the Serre Lab at Brown University in 2011. This dataset consists of 51 action classes inside 6849 clips mostly collected from movies or public video sources. Some exemplary frames used in the HMDB-51 dataset are shown in Figure 2.1. In this dataset there are five groups of actions. First group is general facial actions such as simling, laughing and talking.Second group is facial actions that contain an object such as smoking, eating and drinking. Third group is general body movements such as hand clapping, climbing, diving, flipping, handstanding and jumping. Fourth group is body movements involving interaction with an object such as catching, sword drawing, dribbling, golfing and pushing. Fifth group is body movements that involve interaction with a human such as fencing, hugging, kicking and kissing.

2.5.2.3. JHMDB Dataset. The JHMDB dataset [111] is both a subset and an extension of the HMDB dataset. The JHMDB dataset is generated by selecting subset of the actions in the JHMDB dataset and annotating the human joint positions in the clips. As such, this extension allows the evaluation of the performance of optical flow algorithms in capturing human motion content in clips. Authors of this dataset state that the primary purpose of this extension is to better understand what part of the

Figure 2.2. Screenshots of the 11 actions of UCF YouTube Action Dataset.

human action recognition algorithms need to be developed in order to obtain better results.

2.5.3. Internet Datasets

2.5.3.1. UCF YouTube Action Dataset. This dataset was created in 2009 with videos from YouTube. It contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This dataset is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. For each category, the videos are grouped into 25 groups with more than 4 action clips in it. The video clips in the same group may share some common features, such as the same actor, similar background, similar viewpoint, and so on. The ground-truth is provided in VIPER format giving bounding boxes and action annotation. Figure 2.2 shows two example frames for the 11 actions.

2.5.3.2. ActivityNet Dataset. ActivityNet is a new large-scale video benchmark for human activity understanding. ActivityNet aims at covering a wide range of complex human activities that are of interest to people in their daily living. In its current version, ActivityNet provides samples from 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video, for a total of 849 video hours.

2.5.4. RGB-D Datasets

2.5.4.1. CAD-60 and CAD-120. CAD-60 and its variant CAD-120 [112] are RGB-D datasets recorded using the Microsoft Kinect sensor. RGB-D datasets are more complex as it also gives the depth information. These datasets are commonly used to build 3D body models. The Cad-60 dataset contains 12 actions in 5 different environments performed by 4 different subjects. Action classes are insing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, opening pill container, cooking (chopping), cooking (stirring), talking on couch, relaxing on couch, writing on whiteboard, working on computer. The CAD-120 dataset is more complex since it records longer daily activities that contain sub-activities. There are 10 high level activities and 10 sublevel activities. The high level activities are making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, picking objects, cleaning objects, taking food, arranging objects, having a meal; sublevel activities are reaching, moving, pouring, eating, drinking, opening, placing, closing, scrubbing and null. The CAD-120 dataset also contains affordance labels for the objects in the frames. The affordance labels are reachable, movable, pourable, pourto, containable, drinkable, openable, placeable, closable, scrubbable, scrubber, stationary. Both datasets also contain the traced skeletons.

3. ACTION LOCALIZATION via DEEP TEMPORAL ACTION PROPOSALS

Temporal action localization is the problem of detecting the temporal boundaries which contain a singular action instance in a given video. This is an important first step for an accurate action recognition module. However, this is a challenging task due to the cluttered nature of many real-world action videos. Many of these videos contain complex interaction with other agents, overlapping or irrelevant video segments, occlusions and subtle motion changes.

3.1. Problem Definition

The problem of temporal action localization can be described as follows; in a long and untrimmed video, detect the temporal intervals in which a singular action or activity occurs. Such a video can be described as as a set of frames as follows:

$$X = \{x_n\}_{n=1}^{N} \tag{3.1}$$

where N is the number of frames and x_n is the nth RGB frame. The temporal action localization problem is to find a set of time pairs as follows:

$$\tau_p = \{\phi_n = (t_{s,n}, t_{e,n})\}_{n=1}^{N_p} \tag{3.2}$$

where ϕ_n is the nth time pair, $t_{s,n}$ is the starting time for the nth temporal annotation, ϕ_n is the nth temporal annotation, $t_{e,n}$ is the ending time for the nth temporal annotation and N_p is the number of temporal annotations.

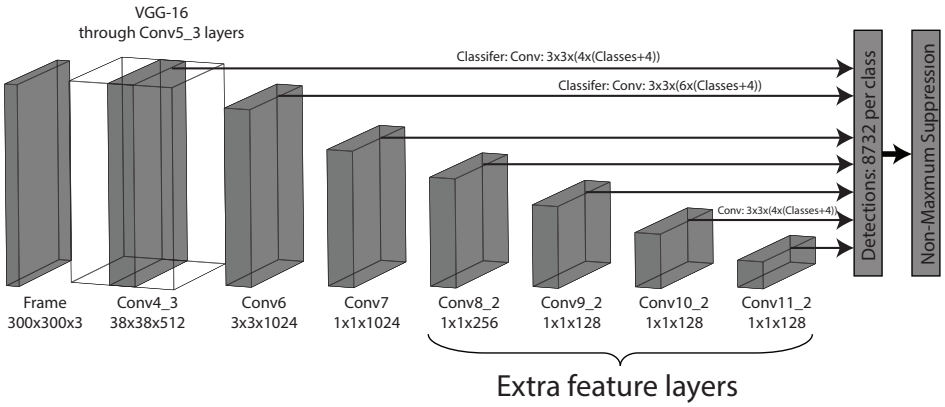

Figure 3.1. The SSD network. The SSD architecture is based on a feed-forward convolutional neural network which provides bounding boxes and scores for object class instances followed by a non-maximum suppression step.

3.2. Action Localization in a Real-Time and Online Setting

The problem of action localization is rendered harder when the action localization is to be done in real-time and in an online setting. Many state-of-the-art techniques work with the assumption that the video that contains the action sequence is available ahead of time. In these settings the action representations need to be properly modified so as to allow for constructing action tubes from an online video source.

3.2.1. Single Shot Detector Model

To achieve action localization at real-time the single shot detector (SSD) model that is described in [113] is employed. The SSD architecture is based on a feed-forward convolutional neural network which provides bounding boxes and scores for object class instances followed by a non-maximum suppression step. A schematic of the SSD network is shown in Figure 3.1. The first layers are designed for the purpose of image classification based on standart convolutional neural networks. These layers are called the base network in the original work by Liu et al. in [113]. The following layers function as a detector network for multi-scale meature maps. These layers get smaller at each step so as to allow detections at multiple scales.

3.2.2. Real-Time Optical Flow Computation

The optical flow methodology used in this thesis is based on the deep inverse search method proposed by Kroeger et al. in [114]. The reason for this selection is the ability of this algorithm to produce optical flow maps at very high frame rates with limited computational resources. The main procedure in this method is an efficient search algorithm for finding patch correspondences between two frames. The search algorithm is based on the gradient descent method where a warping vector $u = (u, v)$ is defined as follows,

$$u = argmin_{u'} \sum [I_{t+1}(x + u') - T(x)]^2. \tag{3.3}$$

This quantity in optimized iteratively using the inverse Lukas-Kanade method [79]. For this we define an update vector Δu such that.

$$\Delta u = argmin_{\Delta u'} \sum_{x} [I_{t+1}(x + u + \Delta u') - T(x)]^2 \tag{3.4}$$

At each level the patches are re-initialized by computing a dense flow field. The dense flow field is computed by the algorithm described next. First the patches are initialized in a uniform grid over the image domain. At the first iteration all patches are intialized with zero flow.

$$\mathbf{u}_{i,init} = \mathbf{U}_{s+1}(\mathbf{x}/\theta_{sd}) \cdot \theta_{sd} \tag{3.5}$$

After the displacement vectors are updated, for all patches that the value $|\mathbf{u}_{i,init} - \mathbf{u}_i|_2$ exceeds the patch size θ_{ps}. After this filtering, dense flow field is calculated by a weighted averaging calculation:

$$U_s(x) = \frac{1}{Z} \sum_{i}^{N_s} \frac{\lambda_{i,x}}{max(1, |d_i(x)|_2)} \cdot u_i. \tag{3.6}$$

A simplified variant of the variational refinement of without a feature matching term and intensity images only is used to refine only on the current scale. The energy is a weighted sum of intensity and gradient data terms (E_I, E_G) and a smoothness term (E_S) over the image domain

$$E(U) = \int_{\Omega} \alpha\Psi(E_I) + \beta\Psi(E_G) + \gamma\Psi(E_S)dx \tag{3.7}$$

This method lends itself to some extensions as follows as described in [114]. This algorithm is easily parallelizable particularly the parts that are computationally intensive. Patch optimization part and the variational refinement parts are linear systems that can operate independently. Furthermore, color information, instead of just intensity increases the performance of optical flow methods.

3.2.3. Bounding Box Prediction

The integrated convolution neural network shown in Figure 3.2 is used for generating action proposals and identifying their spatial bounding boxes. The functionalities of this network are three-fold:

- region proposal generation
- bounding box prediction, and
- estimation of class-specific confidence scores for predicted boxes.

The input image size is selected as 300x300 in order to increase the training and test time of the network. After the calculation of bounding boxes with appearance and flow SSD networks, a fusion step is applied. The fusion methodology is as described in [115].

3.3. Experiments and Results

3.3.1. Evaluation Metric

The performance of the network is measured for the tasks of action label prediction and online temporal and spatial action localization. The evaluation metrics are the standard intersection over union (IoU), area under the curve and mean average precision (mAP). Intersection over union calculation is done as follows:

$$\text{IoU} = \frac{A \cap B}{A \cup B} \tag{3.8}$$

where A and B are the predicted and ground-trouth bounding boxes respectively. IoU evaluation is scale independent and thus is a commonly evaluation metric. Area under the curve metric is an extension of the receiver operating characteristic (ROC) analysis and is defined as the area under the ROC curve. The ROC curve is obtained by plotting true positive and false negative rates against each other at different threshold values. Mean average precision is the arithmetic mean of average precision measurements wich can be expressed as follows

$$MAP = \frac{1}{n} \sum_{n} AP_n \tag{3.9}$$

where AP denotes average precision and n denotes the number of measurements.

3.3.2. Experiments

Experiments are performed on HMDB-21, UCF101, THUMOS14 and ActivityNet datasets. Since THUMOS and ActivityNet datasets lack spatial bounding box annotations, I used the whole frame as the spatial bounding box while training. Online

spatiotemporal localization methods such as the one used in this thesis compute action tubes in an incremental fashion. At each time step, only the frames that were processed up until now are known and the intersection over union (IoU) value can be calculated for these frames only.

3.3.3. Extracted Snippets

In this section the snippets extracted from the action proposal network and training results obtained by feeding these snippets into the pretrained action recognition network will be given. For action proposal a pre-trained network is used on the UCF101 dataset.

Some sets of frames from the snippets extracted via DAPs on the UCF101 dataset are given in Figure 3.3. These snippets are used instead of the randomly selected snippets of TSNs. The obtained snippets are highly conformant with the ground-truth. The network can generate segments at around 121 FPS.

3.3.4. Network Performance

The AUC curves of the network are given in Figure 3.4. It can be seen from this figure that the detection performance increases as a function of the observed number of frames. In addition to this result, in Table 3.1 and Table 3.2 the global detection results for UCF101 and HDMB-21 datasets are given. The AUC curves of the network are given in Figure 3.4. It can be seen from this figure that the detection performance increases as a function of the observed number of frames. In addition to this result, in Table 3.1 and 3.2 the global detection results for UCF101 and HDMB-21 datasets are given. The AUC curves of the network are given in Figure 3.4. It can be seen from this figure that the detection performance increases as a function of the observed number of frames. In addition to this result, in Table 3.1 and Table 3.2 the global detection results for UCF101 and HDMB-21 datasets are given.

Table 3.1. IoU comparison with state-of-the-art methods on the HMDB-51 dataset.

IoU threshold	**0.2**	**0.5**	**0.75**	**0.5:0.95**
Yu et al. [116]	26.5%	-	-	-
Weinzaepfel et al. [117]	46.8%	-	-	-
Peng and Schmid [118]	73.5%	32.1%	73.5%	88.6%
Saha et al. [115]	66.6%	36.4%	66.6%	88.1%
Singh et al. [119]	69.7%	41.9%	14.1%	18.4%
Our results	67.8%	40.8%	18.8%	16.1%

Table 3.2. IoU comparison with state-of-the-art methods on the THUMOS dataset.

IoU threshold	**0.2**	**0.5**	**0.75**	**0.5:0.95**
Gkioxari and Malik [10]	-	53.3	-	-
Wang et al. [120]	46.8%	-	56.4	-
Weinzaepfel et al. [117]	63.1%	60.7%	-	-
Saha et al. [115]	72.6%	71.4%	43.3%	40.0%
Peng and Schmid [118]	74.1%	73.1%	-%	-%
Singh et al. [119]	66.0%	63.9%	35.1%	34.4%
Our results	62.4%	64.5%	38.8%	32.1%

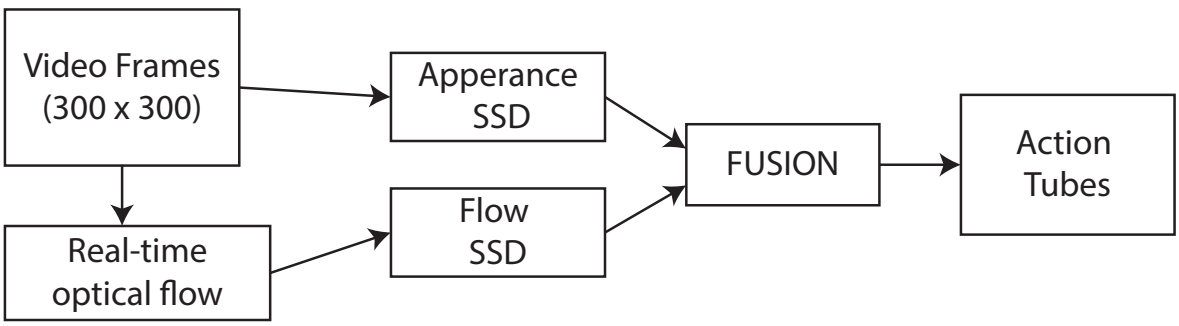

Figure 3.2. Two stream action detection network is based on integrating the single shot detector network architecture into a two-stream convolutional neural network architecture. As such, both the appearance and motion content are considered and fused together to obtain highly confident temporal and spatial action proposals.

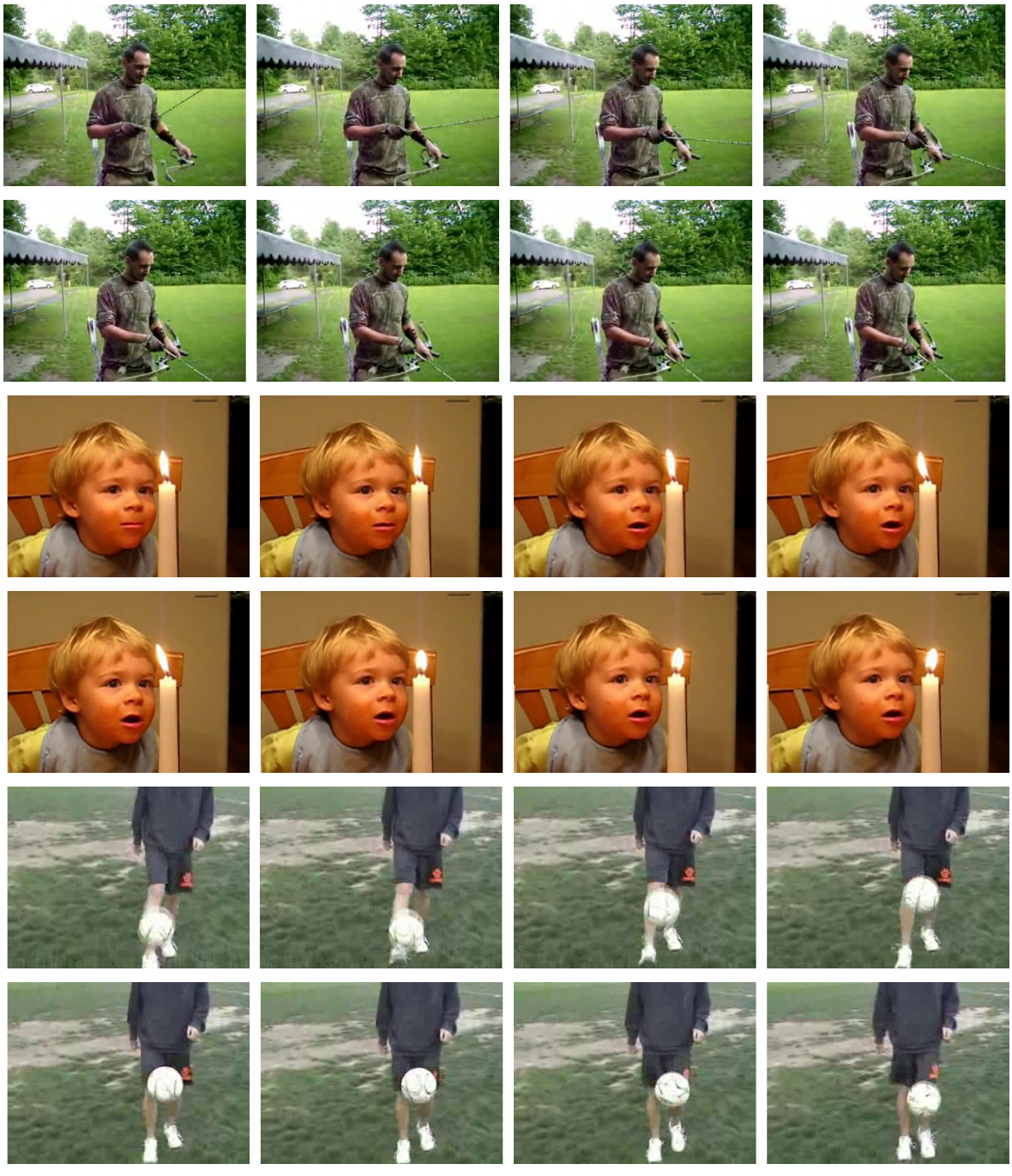

Figure 3.3. Exemplary snippets that were extracted by the method proposed in this papers. Using these snippets instead of uniformly sampled snippets increases the performance of action Temporal Segment Network performance.

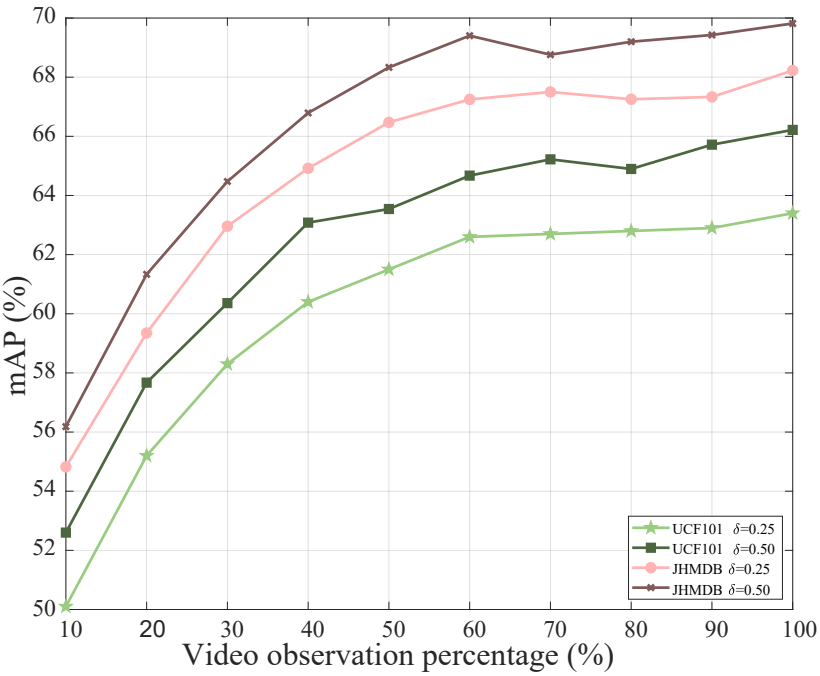

Figure 3.4. Action localisation results using the mAP (%) metric on UCF101-24 and JHMDB, at IoU thresholds of $\delta = 0.25, 0.5$.

4. TEMPORAL SEGMENT NETWORKS for ACTION RECOGNITION

4.1. Two Stream Convolutional Neural Networks

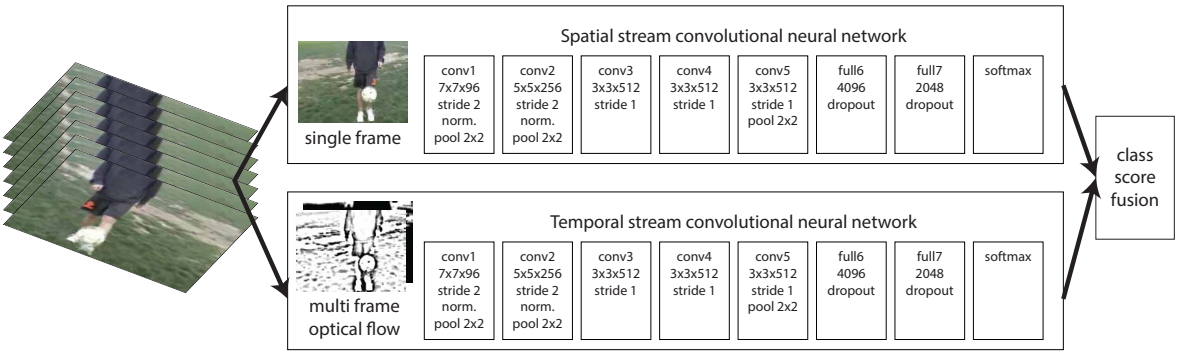

Figure 4.1. Two-stream convolutional neural network architecture used for action recognition. The idea behind this network comes from neurological studies where it was discovered that human brain contained two separate pathways for object recognition and motion recognition.

A figure depicting the two-stream architecture used in [121] is given in Figure 4.1. As can be seen from this figure, the appearance based stream is trained on singular frames from training videos and the motion based stream is trained on a multi-frame optical flow extracted from training videos.

The shortcoming of this architecture is that, it is unable to model the long-term structures present in the video since it is designed to work on short frame stack (e.g., 16 frames) with limited temporal durations. Long term temporal structure is important in the recognition of some actions. Because, complex actions are composed of several motions over a long time period and their sequential relation is an important factor in their definition. Convolutional neural networks lack the ability to model temporal structures.

Wang et al. [98] proposed a novel approach to this problem. In this study, the temporal structure of a given video is captured by training multiple two-stream con-

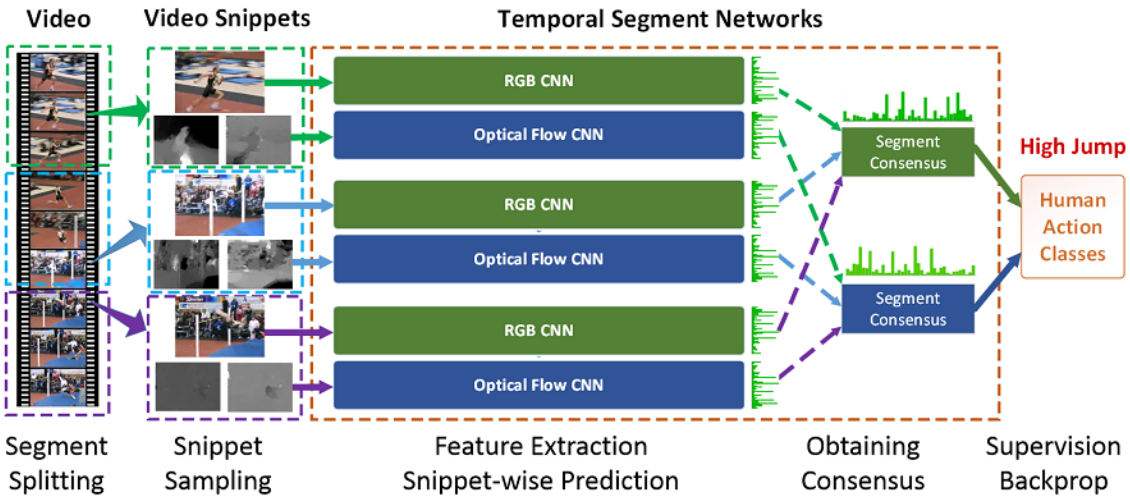

Figure 4.2. Temporal segment network acrhitecture. The two-stream network is parallelized for randomly selected snippets from the whole video. This way temporal structured in the motion and appearence content can be recognized.

volutional networks using short snippets that are sampled from the whole video. The network architecture used in this study is given in Figure 4.2. As can be seen from this figure, for each snippet a two-stream network is trained and the outputs of these networks are fused using a segmental consensus layer for both spatial and temporal networks. Afterwards, class scores from spatial and temporal parts are fused into a single vector of class scores.

4.2. Temporal Structure Modeling

In this section a temporal model that can be used for action recognition will be described. This model is derived from the temporal model that was proposed by Niebles et al. in [122]. The idea is to decomposed a video into viarious temporal segments and match each segment to a motion classifier in accordance with its image-based similarities.

4.3. Temporal Segment Network Architecture

Sampling of snippets from video is done by dividing it into K segments and randomly choosing a short snippet for each segment. This sparse sampling strategy gives an overall idea about the sequential appearance and motion content of the whole video. More formally, given a video V, it is divided into K segments $\{S_1, S_2, \ldots, S_K\}$ of equal duration. Then, the temporal segment network can be expressed as a model of multiple segments as

$$TSN(T_1, T_2, \cdots, T_k) = H(G(F(T_1; \mathbf{W}), F(T_2; \mathbf{W}), \cdots, F(T_K; \mathbf{W}))) \tag{4.1}$$

Here $T_1, T_2, \cdots, T_K$ is the sequence of snippets extracted from video where each snippet is randomly sampled from its corresponding segment S_k. $F(T_k; \mathbf{W})$ is the function representing the convolutional network with parameters $\mathbf{W}$ which operates on the snippet T_k, G is the segmental consensus function which combines the output from multiple snippets to obtain a class hypothesis among them and H is the prediction function which predicts the probability of each action class for the whole video. The prediction function H is chosen to be the widely used Softmax function. Combining with standard categorical cross-entropy loss, the final loss function regarding the segmental consensus is formed as

$$L(y, \mathbf{G}) = -\sum_{i=1}^{G} y_i \left(G_i - \log \sum_{j=1}^{C} \exp G_j \right) \tag{4.2}$$

where C is the number of action classes and y_i is the ground-truth label concerning class i.

This temporal segment network is differentiable or at least has subgradients, depending on the choice of g. This allows us to utilize the multiple snippets to jointly optimize the model parameters $\mathbf{W}$ with standard back-propagation algorithms. In the back-propagation process, the gradients of model parameters $\mathbf{W}$ with respect to the

loss value L can be derived as

$$\frac{\partial L(y, \mathbf{G})}{\partial \mathbf{W}} = \frac{\partial L}{\partial \mathbf{G}} \sum_{k-1}^{K} \frac{\partial G}{\partial F(T_k)} \frac{\partial F(T_k)}{\partial \mathbf{W}} \tag{4.3}$$

where K is the number of segments in the network. In (4.3), it can be seen that the parameter updates of a learning system will utilize the segmental consensus variable $\mathbf{G}$ which is derived from all snippets. This way, the network can model the temporal relationship and learn parameters from the entire video rather than a single snippet.

4.3.1. Training of TSNs

In this part the training process of TSNs will be explained. The TSN network uses the Batch-normalized Inception architecture as a building block [123]. This deep architecture allows modeling and representation of complex actions. In a TSN architecture, the Inception model is adapted as a two-stream network model in which one of the streams try to capture motion information from the video. However, as the authors of [98] point out, different input modalities can be used such as the warped optical flow which might achieve better performance in suppressing the effect of camera motion.

4.3.2. Testing of TSNs

In this part the testing procedure for the TSNs is explained. During the testing we perform frame-wise evaluation on the action proposals obtained by the action proposal network detailed in Chapter 3. In the original work the snippets are of fixed-length, uniformly sampled throughout the video. In this thesis, the time warped segments are used in the testing procedure. Both the RGB data and flow data obtained from the real-time optical flow algorithm described in Subsection 3.2.2. The fusion of spatial temporal information is done by taking a weighted average of them. The warped optical flow modality is not used during the testing.

5. TRANSFER LEARNING of TSNs via ACTION PROPOSALS

5.1. Retraining the Output Layer

One of the most commonly used transfer learning methods in machine learning methodology is reinitializing and retraining the output layer of the network. The output layer is retrained by optimizing the cross entropy loss over posterior probabilities over the ground-truth action class.

$$p(c_t|x_{t-\delta}^{t+\delta}) = p(c_t|x_t) = \text{softmax}(\mathbf{W}^T x_t + b) \tag{5.1}$$

where c_t is the ground-truth class label over the window $(t-\delta, t+\delta)$. A drawback of this methodology is that only a linear transformation can be trained from the features of the pre-trained network over the new dataset. However, with enough training data this methodology produces satisfactory results.

5.2. Experiments and Results

The results on the test videos are summarized in Table 5.1. The performance of the network is decreased relative to the original TSN network with an accuracy of 93.8% against an accuracy of 94.9% for the UCF101 dataset, an accuracy of 70.5% against an accuracy of 71.0% for the HMDB51 dataset, an accuracy of 80.1% against an accuracy of 89.6% for the THUMOS14 dataset and an accuracy of 79.4% against an accuracy of 85.2% for the ActivityNet dataset. This shows that the obtained results are comparable with that of the state-of-the-art methods.

Table 5.1. Performance comparison of the proposed TSN+AP network with current state-of-the-art networks [98].

	HMDB51	**UCF101**	**THUMOS14**	**ActivityNet**
TwoStream	59.4%	88.0%	66.1%	71.9%
VideoDarwin	63.7%	85.2%	-	-
MPR	65.5%	88.6%	-	-
F_{ST}CN	59.1%	88.1%	-	-
TDD+FV	63.2%	90.3%	-	-
LTC	64.8%	91.7%	-	-
KVMF	63.3%	93.1%	-	-
iDT+FV	-	-	63.1%	66.5%
object+motion	-	-	71.6%	78.1%
EMV+RGB	-	-	61.5%	74.1%
TSN	71.0%	94.9%	80.1%	89.6%
TSN+AP	70.5%	93.8%	79.4%	85.2%

6. CONCLUSION

In this work we presented a novel framework for strategically sampling a video stream for robustly detecting segments that ave a high actionness score and training a temporal segment network using these segments. Current results of the network yields a comparable result with the state-of-the-art methods. However, the need for extracting action proposals ahead of time also adds a level of complexity and computational cost. In order to compensate for this added level of complexity, it might be worthy of consideration to use a network that is trained with snippets extracted by action proposal networks. Overall, developing better strategies for temporally sampling a video stream increases the training performance and also provides decreases the computational burden particularly in real-time applications.

REFERENCES

1. Weinland, D., R. Ronfard and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition", *Computer Vision and Image Understanding*, Vol. 115, No. 2, pp. 224–241, 2011.

2. Wang, L., Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition", *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 20–36, 2016.

3. Turaga, P., R. Chellappa, V. S. Subrahmanian and O. Udrea, "Machine recognition of human activities: A survey", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 18, No. 11, pp. 1473–1488, 2008.

4. Poppe, R., "A survey on vision-based human action recognition", *Image and Vision Computing*, Vol. 28, No. 6, pp. 976–990, 2010.

5. Chaaraoui, A. A., P. Climent-Pérez and F. Flórez-Revuelta, "A review on vision techniques applied to Human Behaviour Analysis for Ambient-Assisted Living", *Expert Systems with Applications*, Vol. 39, No. 12, pp. 10873–10888, 2012.

6. Chaquet, J. M., E. J. Carmona and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition", *Computer Vision and Image Understanding*, Vol. 117, No. 6, pp. 633–659, 2013.

7. Wang, X., A. Farhadi and A. Gupta, "Actions ˜ Transformations", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2658–2667, 2016.

8. Moeslund, T. B., A. Hilton and V. Krüger, "A survey of advances in vision-based human motion capture and analysis", *Computer Vision and Image Understand-*

ing, Vol. 104, No. 2, pp. 90–126, 2006.

9. Schaal, S., A. Ijspeert and A. Billard, "Computational approaches to motor learning by imitation", *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, Vol. 358, No. 1431, pp. 537–547, 2003.

10. Laptev, I. and T. Lindeberg, "Space-time interest points", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 432–439, 2003.

11. Gkioxari, G. and J. Malik, "Finding action tubes", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 759–768, 2015.

12. Dautenhahn, K. and A. Billard, "Bringing up robots or the psychology of socially intelligent robots", *Proceedings of the 3rd Annual Conference on Autonomous Agents (AGENTS)*, pp. 366–367, 1999.

13. Breazeal, C., "Toward sociable robots", *Robotics and Autonomous Systems*, Vol. 42, No. 3, pp. 167 – 175, 2003.

14. Fong, T., I. Nourbakhsh and K. Dautenhahn, "A survey of socially interactive robots", Vol. 42, No. 3-4, pp. 143–166, 2003.

15. Mori, M., K. F. MacDorman and N. Kageki, "The Uncanny Valley From the Field", *IEEE Robotics Automation Magazine*, Vol. 19, No. 2, pp. 98–100, 2012.

16. Chrungoo, A., S. S. Manimaran and B. Ravindran, "Activity Recognition for Natural Human Robot Interaction", *Proceedings of the International Conference on Social Robotics (ICSR)*, pp. 84–94, 2014.

17. Coppola, C., S. Cosar, D. R. Faria and N. Bellotto, "Social Activity Recognition on Continuous RGB-D Video Sequences", *International Journal of Social Robotics*, Vol. 12, No. 1, pp. 201–215, 2020.

18. Villani, V., F. Pini, F. Leali and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications", *Mechatronics*, Vol. 55, pp. 248–266, 2018.

19. Liu, Y. and G. Nejat, "Robotic urban search and rescue: A survey from the control perspective", *Journal of Intelligent and Robotic Systems*, Vol. 72, No. 2, pp. 147–165, 2013.

20. Vasconez, J. P., G. A. Kantor and F. A. Auat Cheein, "Human–robot interaction in agriculture: A survey and current challenges", *Biosystems Engineering*, Vol. 179, pp. 35–48, 2019.

21. Beasley, R. A., "Medical Robots: Current Systems and Research Directions", *Journal of Robotics*, Vol. 2012, pp. 1–14, 2012.

22. Patil, D., M. Ansari, D. Tendulkar, R. Bhatlekar, V. N. Pawar and S. Aswale, "A Survey on Autonomous Military Service Robot", *Proceedings of the International Conference on Emerging Trends in Information Technology and Engineering (ETITE)*, pp. 1–7, 2020.

23. Caba Heilbron, F., V. Escorcia, B. Ghanem and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015.

24. Goodale, M. A. and A. Milner, "Separate visual pathways for perception and action", *Trends in Neurosciences*, Vol. 15, No. 1, pp. 20–25, 1992.

25. Caba Heilbron, F., J. Carlos Niebles and B. Ghanem, "Fast temporal activity proposals for efficient detection of human actions in untrimmed videos", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1914–1923, 2016.

26. Mettes, P., J. C. Van Gemert, S. Cappallo, T. Mensink and C. G. Snoek, "Bag-of-fragments: Selecting and encoding video fragments for event detection and recounting", *Proceedings of the ACM Conference on International Conference on Multimedia Retrieval (ICMR)*, pp. 427–434, 2015.

27. Shou, Z., D. Wang and S. Chang, "Action temporal localization in untrimmed videos via multi-stage CNNs", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1049–1058, 2016.

28. Escorcia, V., F. Caba Heilbron, J. C. Niebles and B. Ghanem, "DAPs: Deep Action Proposals for Action Understanding", *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 768–784, 2016.

29. Buch, S., V. Escorcia, C. Shen, B. Ghanem and J. C. Niebles, "SST: Single-Stream Temporal Action Proposals", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6373–6382, 2017.

30. Gao, J., K. Chen and R. Nevatia, "Ctap: Complementary temporal action proposal generation", *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 68–83, 2018.

31. Wang, H. and C. Schmid, "Action recognition with improved trajectories", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3551–3558, 2013.

32. Burghouts, G. J., K. Schutte, R. J. ten Hove, S. P. van den Broek, J. Baan, O. Rajadell, J. R. van Huis, J. van Rest, P. Hanckmann, H. Bouma, G. Sanroma, M. Evans and J. Ferryman, "Instantaneous threat detection based on a semantic representation of activities, zones and trajectories", *Signal, Image and Video Processing*, Vol. 8, No. 1, pp. 191–200, 2014.

33. Oreifej, O. and Z. Liu, "HON4D: Histogram of oriented 4D normals for activ-

ity recognition from depth sequences", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 716–723, 2013.

34. Yang, X. and Y. Tian, "Effective 3D action recognition using EigenJoints", *Journal of Visual Communication and Image Representation*, Vol. 25, No. 1, pp. 2–11, 2014.

35. Ye, M., Q. Zhang, L. Wang, J. Zhu, R. Yang and J. Gall, "A survey on human motion analysis from depth data", *Lecture Notes in Computer Science*, Vol. 8200, pp. 149–187, 2013.

36. Han, F., B. Reily, W. Hoff and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review", *Computer Vision and Image Understanding*, Vol. 158, pp. 85–105, 2017.

37. Johansson, G., "Visual perception of biological motion and a model for its analysis", *Perception & Psychophysics*, Vol. 14, No. 2, pp. 201–211, 1973.

38. Cutting, J. E. and L. T. Kozlowski, "Recognizing friends by their walk: Gait perception without familiarity cues", *Bulletin of the Psychonomic Society*, Vol. 9, No. 5, pp. 353–356, 1977.

39. Dittrich, W. H., "Action categories and the perception of biological motion", *Perception*, Vol. 22, No. 1, pp. 15–22, 1993.

40. Herath, S., M. Harandi and F. Porikli, "Going Deeper into Action Recognition: A Survey", *Image and Vision Computing*, Vol. 60, pp. 4–21, 2016.

41. Ramanan, D. and D. A. Forsyth, "Automatic Annotation of Everyday Movements", *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1547–1554, 2004.

42. Green, R. D. and L. Guan, "Quantifying and recognizing human movement pat-

terns from monocular video images - Part II applications to biometrics", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 14, No. 2, pp. 191–198, 2004.

43. Lv, F. and R. Nevatia, "Recognition and segmentation of 3-D human action using HMM and multi-class AdaBoost", *Lecture Notes in Computer Science*, Vol. 3954, pp. 359–372, 2006.

44. Wang, Y., H. Jiang, M. S. Drew, Z. N. Li and G. Mori, "Unsupervised discovery of action classes", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2, pp. 1654–1661, 2006.

45. Guerra-Filho, G. and Y. Aloimonos, "A language for human action", *Computer*, Vol. 40, No. 5, pp. 42–51, 2007.

46. Gavrila, D. M. and L. S. Davis, "3-D model-based tracking of humans in action: a multi-view approach", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 73–80, 1996.

47. Yacoob, Y. and M. J. Black, "Parameterized Modeling and Recognition of Activities", *Computer Vision and Image Understanding*, Vol. 73, No. 2, pp. 232–247, 1999.

48. Rao, C., A. Yilmaz and M. Shah, "View-invariant representation and recognition of actions", *International Journal of Computer Vision*, Vol. 50, No. 2, pp. 203–226, 2002.

49. Ikizler, N. and P. Duygulu, "Histogram of oriented rectangles: A new pose descriptor for human action recognition", *Image and Vision Computing*, Vol. 27, No. 10, pp. 1515–1526, 2009.

50. Elgammal, A., V. Shet, Y. Yacoob and L. S. Davis, "Learning dynamics for exemplar-based gesture recognition", *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), Vol. 1, pp. 1–5, 2003.

51. Yuping Shen and H. Foroosh, "View-invariant recognition of body pose from space-time templates", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–6, 2008.

52. Turaga, P., A. Veeraraghavan and R. Chellappa, "Statistical analysis on stiefel and grassmann manifolds with applications in computer vision", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

53. Weinland, D., E. Boyer and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–7, 2007.

54. Lv, F. and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2007.

55. Shi, Q., L. Wang, L. Cheng and A. Smola, "Discriminative human action segmentation and recognition using semi-markov model", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

56. Natarajan, P. and R. Nevatia, "View and scale invariant action recognition using multiview shape-flow models", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

57. Bobick, A. F. and J. W. Davis, "The recognition of human movement using temporal templates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, pp. 257–267, 2001.

58. Weinland, D., R. Ronfard and E. Boyer, "Free viewpoint action recognition using motion history volumes", *Computer Vision and Image Understanding*, Vol. 104,

No. 2, pp. 249–257, 2006.

59. Laptev, I. and P. Perez, "Retrieving actions in movies", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

60. Fathi, A. and G. Mori, "Action recognition by learning mid-level motion features", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–7, 2008.

61. Souvenir, R. and J. Babbs, "Learning the viewpoint manifold for action recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–7, 2008.

62. Farhadi, A. and M. K. Tabrizi, "Learning to Recognize Activities from the Wrong View Point", *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 154–166, 2008.

63. Efros, A. A., A. C. Berg, G. Mori and J. Malik, "Recognizing action at a distance", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 726–733, 2003.

64. Weinland, D. and E. Boyer, "Action recognition using exemplar-based embedding", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–7, 2008.

65. Schindler, K. and L. Van Gool, "Action Snippets: How many frames does human action recognition require?", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

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

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

68. Ke, Y., R. Sukthankar and M. Hebert, "Event detection in crowded videos", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

69. Schuldt, C., I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach", *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 32–36, 2004.

70. Boiman, O. and M. Irani, "Detecting irregularities in images and in video", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Vol. I, pp. 462–469, 2005.

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

72. Niebles, J. C., 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.

73. Niebles, J. C. and F. F. Li, "A hierarchical model of shape and appearance for human action classification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2007.

74. Klaeser, A., M. Marszalek and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients", *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1–10, 2008.

75. Bobick, A. F. and J. W. Davis, "The recognition of human movement using temporal templates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, pp. 257–267, 2001.

76. Hu, Y., L. Cao, F. Lv, S. Yan, Y. Gong and T. S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 128–135, 2009.

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

78. Wang, H., A. Kläser, C. Schmid and C.-L. Liu, "Action recognition by dense trajectories", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3169–3176, 2011.

79. Johansson, G., "Visual motion perception", *Scientific American*, Vol. 232, No. 6, pp. 76–89, 1975.

80. Lucas, B. D., T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision", *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 674–679, 1981.

81. Darrell, T. and A. Pentland, "Space-time gestures", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 335–340, 1993.

82. Gavrila, D. M. and L. S. Davis, "Towards 3-D model-based tracking and recognition of human movement: a multi-view approach", *Proceedings of the IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 272–277, 1995.

83. Yamato, J., J. Ohya and K. Ishii, "Recognizing human action in time-sequential

images using hidden markov model", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 379–385, 1992.

84. Gong, S. and T. Xiang, "Recognition of group activities using dynamic probabilistic networks", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 742–749, 2003.

85. Han, L., X. Wu, W. Liang, G. Hou and Y. Jia, "Discriminative human action recognition in the learned hierarchical manifold space", *Image and Vision Computing*, Vol. 28, No. 5, pp. 836–849, 2010.

86. Intille, S. S. and A. F. Bobick, "A framework for recognizing multi-agent action from visual evidence", *AAAI-99 Proceedings*, pp. 518–525, 1999.

87. Ghanem, N., D. DeMenthon, D. Doermann and L. Davis, "Representation and recognition of events in surveillance video using petri nets", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 112–113, 2004.

88. Lowe, D. G., "Object recognition from local scale-invariant features", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Vol. 2, pp. 1150–1157, 1999.

89. Skrypnyk, I. and D. G. Lowe, "Scene modelling, recognition and tracking with invariant image features", *Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 110–119, 2004.

90. Scovanner, P., S. Ali and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition", *Proceedings of the ACM International Multimedia Conference and Exhibition (ICME)*, pp. 357–360, 2007.

91. Al Ghamdi, M., L. Zhang and Y. Gotoh, "Spatio-Temporal SIFT and Its Application to Human Action Classification", *Proceedings of the IEEE International*

Conference on Computer Vision (ICCV), pp. 301—310, 2012.

92. Willems, G., T. Tuytelaars and L. Van Gool, "An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector", *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 650–663, 2008.

93. Ji, S., W. Xu, M. Yang and K. Yu, "3D Convolutional neural networks for human action recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 1, pp. 221–231, 2013.

94. McIntosh, R. D. and T. Schenk, "Two visual streams for perception and action: Current trends", *Neuropsychologia*, Vol. 47, No. 6, pp. 1391–1396, 2009.

95. Simonyan, K. and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", *Advances in Neural Information Processing Systems*, Vol. 1, No. 1, pp. 568–576, 2014.

96. Feichtenhofer, C., A. Pinz and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1933–1941, 2016.

97. Ahsan, U., C. Sun and I. Essa, "DiscrimNet: Semi-Supervised Action Recognition from Videos using Generative Adversarial Networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10, 2018.

98. Wang, L., C. Gao, L. Yang, Y. Zhao, W. Zuo and D. Meng, "PM-GANs: Discriminative Representation Learning for Action Recognition Using Partial Modalities", *Lecture Notes in Computer Science*, Vol. 11210, pp. 389–406, 2018.

99. Soomro, K., A. R. Zamir and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–6, 2012.

100. Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1725–1732, 2014.

101. Marszalek, M., I. Laptev and C. Schmid, "Actions in context", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2929–2936, 2009.

102. Sigal, L., A. O. Balan and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion", *International Journal of Computer Vision*, Vol. 87, No. 1, pp. 4–27, 2010.

103. Gorban, A., H. Idrees, Y. Jiang, A. R. Zamir, I. Laptev, M. Shah and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–6, 2015.

104. Wang, J., Z. Liu, Y. Wu and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1297, 2012.

105. Koppula, H. S., R. Gupta and A. Saxena, "Learning human activities and object affordances from RGB-D videos", *International Journal of Robotics Research*, Vol. 32, No. 8, pp. 951–970, 2013.

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

107. Moussa, M. M., E. Hamayed, M. B. Fayek and H. A. El Nemr, "An enhanced method for human action recognition", *Journal of Advanced Research*, Vol. 6, No. 2, pp. 163–169, 2015.

108. Rahman, S. A., S. Y. Cho and M. K. Leung, "Recognising human actions by analysing negative spaces", *IET Computer Vision*, Vol. 6, No. 3, pp. 197–213, 2012.

109. Vishwakarma, D. K. and R. Kapoor, "Hybrid classifier based human activity recognition using the silhouette and cells", *Expert Systems with Applications*, Vol. 42, No. 20, pp. 6957–6965, 2015.

110. Yuan, J., Z. Liu and Y. Wu, "Discriminative video pattern search for efficient action detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 9, pp. 1728–1743, 2011.

111. Jhuang, H., J. Gall, S. Zuffi, C. Schmid and M. J. Black, "Towards understanding action recognition", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3192–3199, 2013.

112. Sung, J., C. Ponce, B. Selman and A. Saxena, "Unstructured human activity detection from RGB-D images", *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 842–849, 2012.

113. Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, "SSD: Single Shot MultiBox Detector", *Lecture Notes in Computer Science*, Vol. 9905, pp. 21–37, 2015.

114. Kroeger, T., R. Timofte, D. Dai and L. Van Gool, "Fast Optical Flow using Dense Inverse Search", *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 1–20, 2016.

115. Saha, S., G. Singh, M. Sapienza, P. H. S. Torr and F. Cuzzolin, "Deep Learning

for Detecting Multiple Space-Time Action Tubes in Videos", *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1–10, 2016.

116. Yu, G. and J. Yuan, "Fast action proposals for human action detection and search", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1302–1311, 2015.

117. Weinzaepfel, P., Z. Harchaoui and C. Schmid, "Learning to Track for Spatio-Temporal Action Localization", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3164–3172, 2015.

118. Peng, X. and C. Schmid, "Multi-region Two-Stream R-CNN for Action Detection", *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 744–759, 2016.

119. Singh, G., S. Saha, M. Sapienza, P. Torr and F. Cuzzolin, "Online Real-time Multiple Spatiotemporal Action Localisation and Prediction", pp. 3657–3666, 2016.

120. Wang, L., Y. Qiao, X. Tang and L. Van Gool, "Actionness Estimation Using Hybrid Fully Convolutional Networks", pp. 2708–2717, 2016.

121. Simonyan, K. and A. Zisserman, "Two-stream convolutional networks for action recognition in videos", *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 568–576, 2014.

122. Niebles, J. C., C. W. Chen and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification", *Lecture Notes in Computer Science*, Vol. 6312, pp. 392–405, 2010.

123. Ioffe, S. and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.