HUMAN ACTIVITY RECOGNITION IN LOW QUALITY VIDEOS USING SPATIO-TEMPORAL FEATURES

SAIMUNUR RAHMAN

MASTER OF SCIENCE
(INFORMATION TECHNOLOGY)

MULTIMEDIA UNIVERSITY

JUNE 2016

HUMAN ACTIVITY RECOGNITION IN LOW QUALITY VIDEOS USING SPATIO-TEMPORAL FEATURES

BY

SAIMUNUR RAHMAN

B.Sc. (Engg.), International Islamic University Chittagong, Bangladesh

THESIS SUBMITTED IN FULFILMENT OF THE

REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

(by Research)

in the

Faculty of Computing and Informatics

MULTIMEDIA UNIVERSITY
MALAYSIA

June 2016

DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this Thesis has been submitted in support of any application for any other degree or qualification on this or any other university or institution of learning.

_____

Saimunur Rahman

ACKNOWLEDGEMENTS

First and above all, I praise Allah (Swt.), the almighty for providing me this opportunity and granting me the capability to proceed successfully. This thesis appears in its current form due to the assistance and guidance of several people. I would therefore like to offer my sincere thanks to all of them.

Second, I want to thank my supervisor Dr. John See. It has been an honor to be his first postgraduate student. He has taught me, both consciously and unconsciously, how good research in computer vision is done. I appreciate all his contributions of time, ideas, moral support and funding to make my master's degree experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the research pursuit. I am also thankful for the excellent example he has provided as a successful computer vision researcher and instructor. I could not have imagined having a better advisor and mentor for my postgraduate studies.

Besides my supervisor, I would like to thank my co-supervisor Dr. Peter Ho for his insightful comments and encouragements. His door was always open whenever I ran into a trouble spot for administrative path holes or had a question about my research or writing.

My time at MMU was made enjoyable in large part due to the many friends that became a part of my life. I am grateful for time spent with friends, our memorable trips into the popular places of West Malaysia, Mohammed and Bahar's hospitality after coming to Malaysia, and for many other people and memories.

A special thanks to my family. Words cannot express how grateful I am with my mother, father and brother for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank all of my friends who supported me, and incented me to strive towards my goal.

To my supervisor *Dr. John See*, my parents, and my brother *Fahad*.

ABSTRACT

Human activity recognition (HAR) is one of the most intensively studied areas of computer vision in recent times. However, under real world conditions, especially when public infrastructure such as surveillance and web cameras are considered, current HAR techniques do not adapt to lower quality videos due to various challenges such as noise and lighting changes, motion blur, poor resolution and sampling.

The objective of this research is to develop a framework and methods for human activity recognition using spatio-temporal information from low quality video. The contributions of this thesis are three-fold. Firstly, a framework based on popular BoVW model for activity recognition from low quality videos is proposed. Secondly, a spatio-temporal joint feature utilization method is proposed to achieve better robustness in the case of low quality videos. The method uses textural features to improve the performance of shape and motion features for better activity recognition in low quality videos. Finally, a spatio-temporal mid-level feature bank (STEM) that encodes visual features into mid-level representation is designed. In STEM, a new approach to spatio-temporal textural feature extraction that extracts discriminate textures from 3-D salient patches is proposed.

Evaluations were conducted on the proposed methods with various low quality versions or subsets of three publicly available datasets: KTH action, UCF-11 and HMDB51. Experimental results shows that proposed methods achieves ≈10-22%, ≈8-10%, ≈10-16%, and ≈13-16% improvement over the baseline results of KTH-SD_4, YouTube-LQ, HMDB51-BQ and HMDB51-MQ low quality versions or subsets. In STEM, the use of salient texture features further improves the recognition performance by considering only the salient part of the video frame. Overall, it can be observed that texture is an important visual feature cue for low quality videos, and the robustness of shape and motion feature can be strengthened by using this.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

HAR	Human Activity Recognition
LQ	Low Quality
MQ	Medium Quality
HQ	High Quality
HD	High Definition
STV	Space-time Volume
STT	Space-time Trajectories
STF	Space-time Features
STIP	Space-time Interest Points
iDT	Improved Dense Trajectories
DT	Dense Trajectories
VQ	Vector Quantization
FV	Fisher Vector
SVM	Support Vector Machine
HOG	Histogram of Oriented Gradients
HOF	Histogram of Optical Flow
MBH	Motion Boundary Histogram
LBP	Local Binary Pattern
LPQ	Local Phase Quantization
BSIF	Binarized Statistical Image features
JFUF	Joint Feature Utilization Framework
STEM	Spatio-temporal Mid-level Feature Bank
KTH Action	Kungliga Tekniska högskolan Human Action Dataset
UCF-11	University Central Florida 11 Class Dataset
HMDB51	Human Motion Database (51 Action Class)
CNN	Convolutional Neural Network

CHAPTER 1

INTRODUCTION

Human activity recognition (HAR) has been an active area of research in the field of computer vision and pattern recognition, producing an enormous amount of progress in recent years. This is partly because of the rapid increase amount of video data capture and the large number of potential application domains based on automated video analysis such as video surveillance, human-computer interaction, sports video analysis, crowd scene analysis, and video indexing and retrieval (Aggarwal & Ryoo, 2011).

HAR from low quality videos such as closed-circuit television cameras (CCTV) and web cameras remains a challenging issue due to various complex problems such as low resolution, low framerate, compression artifacts and motion blur together with general purpose activity recognition problems such as inter and intra class variations, noise and illumination variations, pose and scene variations, motion variation and higher dimensionality. Though much progress have been done for handling many complex video issues such as illumination variations, scale variations and scene variations but problems related with video quality are still considered unexplored (Kuehne, Jhuang, Garrote, Poggio, & Serre, 2011). Recent methods are highly concentrated on high quality videos and does process them tediously with high computational cost for HAR. The complexity burden of current methods are not suitable for the interpretation of human activities in low quality videos.

1.1 Research Overview

With the increasing use of video acquisition devices such as CCTVs, web cameras and mobile devices in our practical life, a large amount of video data is being generated every day. Among the videos generated, humans remain at the center of interest

in most of the videos, and the application of analyzing human activities includes video surveillance, video indexing, human-computer interaction etc. (Aggarwal & Ryoo, 2011; H. Xu et al., 2015). A standard activity recognition system or framework is supposed to recognize the human actions from a video sequence despite of a number of agent interactions captured from an unconstrained environment. Eventhough, much progress have been done in literature, but the recognition of accurate activities from real-world low quality videos still remains a challenging task.

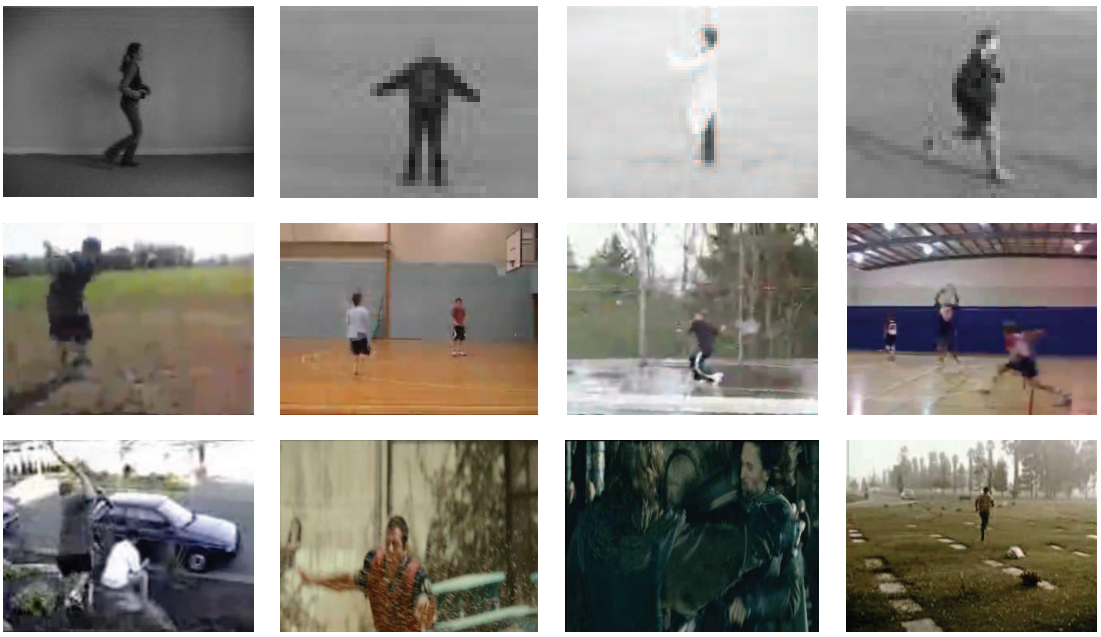


Figure 1.1: Sample low quality videos (resized to same resolution for display) from which we aim to recognize human activities. Samples were taken from KTH action (Schüldt et al., 2004), UCF-11 (J. Liu et al., 2009) and HMDB51 (Kuehne et al., 2011) datasets.

There are a few notable challenges in the HAR task. The first inter and intra class variability of activity classes. A human can perform an activity in various directions and movements, while the activities of two or more class might be separated by small subtle differences. Illumination and viewpoint variations, occlusions, motion variations between same activities also make the feature extraction and classification process challenging. The second is the problem of dimensionality, images are typically dimensions of spatial domains, i.e. heights and widths, while videos also consider the temporal dimensions, i.e. frames. This makes video comparatively higher than the

dimension of image which lead to many computational overhead problems. The final is human activities are generally consists of many sub-activities. Every sub-activity is further decomposed into motion and gestures of various parts of body. Thus, in order to recognize human activities effectively, the framework should have robustness towards in these challenges. Specifically, in feature selection step, it should select efficient features and describe them using effective descriptors. And in classification process, it should have activity specific good numerical models from descriptor feature vectors which turns the contextual information into the correct predictions.

Activity recognition from low quality videos provides a different challenge in addition to the general purpose activity recognition challenges mentioned. Figure 1.1 shows some samples of human actions in low quality videos – videos that are downsampled spatially, compressed, captured with occlusions and irregular camera motion. The videos in the first row are downsampled spatially, the second row videos are compressed and the last row contains videos with occlusions and camera motions. In most of these videos, it is noticeable that the structures of different human body parts are hardly recognizable. It is also hard to differentiate between the background and foreground, human and object. There is no universal rule for measuring video quality, for example, the videos that are with 720p settings (1280$\times$720 resolution) are considered as entry-level high definition (HD) video, but anything less than it is not necessarily low quality. Compression factors such as CRF (constant rate factor) for x264 encoder indirectly controls quality of video encoded, it is still not a measure of quality. HMDB51 (Kuehne et al., 2011) human motion database had videos with 'bad', 'medium', and 'good' quality labels. They determined those labels by subjective evaluations of human experts. The experts gave labels based on the spatial quality of video frames. Since there is no specific measure for 'low video quality' so, video generated or annotated with respect to visual quality in the spatial and temporal domain by all these approximations can be considered as a low quality video. In this research, we intent to recognize human activities from low quality videos.

1.2 Problem Statements and Motivations

In this section, the motivations to pursuit of this area of research from two distinct viewpoints, namely recognition framework and video feature representation perspective are discussed.

1.2.1 Activity Recognition Framework Perspective

From the framework perspective, research in human activity recognition in recent years have focused in many frameworks (X. Wang et al., 2013; Peng et al., 2014). Most of them were designed in such as way that they are only applicable for recognizing human activities in high quality videos. Due to the nature and complexity of low quality video data, existing frameworks are not sufficient enough to process them effectively. They are mostly concentrated on getting the "most" out of good quality videos, by extracting rich feature sets and processing them tediously with high computational costs. There are many issues that contribute to the deterioration of videos such presence of poor weather conditions, camera jitter, motion blurring, low resolution and frame rate, or internal noise. Existing frameworks does not consider these problems as an issue. Owing the drawbacks of existing frameworks, the question comes is - *how do we incorporate the various features from low quality videos in the activity recognition framework to improve the recognition performance?*.

1.2.2 Video Feature Representation

Based on the survey by Aggarwal and Ryoo (2011), there are mainly two types of approaches available in literature, single-layered and hierarchical approaches. The single-layered approaches use video sequences while hierarchical approaches use video sub-events to recognize human activities. However, in terms of popularity, single-layered or spatio-temporal approaches becomes more popular than hierarchical approaches due to their direct action modeling capability from raw video data (Aggarwal & Ryoo, 2011). There is a significant amount of research reported for spatio-temporal approaches in recent years and many offers state-of-the-art activity recognition performance.

Existing spatio-temporal image and video representation methods are not sufficiently robust and effective in discriminating human activities in low quality videos. Many existing methods are carefully hand-crafted to obtain high quality features in simple activities and interactions. In design, additional considerations are also needed to address the deterioration of fidelity, resolution and illumination in low quality videos. Generally, spatio-temporal (or space-time) methods can represent activity in three ways – volumes, trajectories or a set of features (Aggarwal & Ryoo, 2011). Each of these distinct representations have their strengths and drawbacks.

While there is an increase of attention in feature based spatio-temporal approaches due to their reliability under noise and illumination changes, the major limitation lies in its suitability for modeling more complex activities, not just simple periodic activities involving person interactions only. Despite the ability to handle scale invariance well, feature-based approaches struggle to handle viewpoint invariance, especially in wide surveillance areas where persons or objects of interest can undergo a tremendous change of view angle as they move in the monitored scene. On the other hand, trajectory-based spatio-temporal approaches are view-invariant, but this comes at a cost – low-level estimation of 3-D XYZ joint or body part locations of moving persons is a difficult and expensive requirement to success. Moreover, using trajectories for person-object interactions remain a new direction of research. Volume-based approaches are less popular in literature due to its major disadvantage of recognizing actions involving multiple persons. There is an apparent difficulty in spatially segmenting the volume of actions belonging to persons in close proximity. Nevertheless, volume-based approaches offer some high-level representation of human actions (with minimal loss of information) across the temporal dimension, with elegant similarity matching techniques between volumetric patches.

Many of the spatio-temporal feature based approaches use either shape and motion features or both to describe the visual pattern of the action videos. In recent approaches such as space-time interest point (Laptev et al., 2008), dense sampling (H. Wang et al., 2009) and improved dense trajectories (H. Wang & Schmid, 2013),

the use of gradient feature, optical flow and motion boundary features have become popular for their performance. However, most of these features were specifically designed for relatively good quality videos which have high fidelity of signal, detailed spatial resolution, motion consistency. The feature selection and generation strategies of these methods are not always suitable for low quality videos. For instance, the HOG features (Dalal & Triggs, 2005) which are entirely based on the image gradients does not offer a rich set of statistics when quality of the video deteriorates. Figure 1.2 shows the histogram of gradient (HOG) features when videos are compressed. The first image is the reference and second image is its associated HOG vectors, the third and fourth image respectively represents the HOG features when videos are re-encoded, i.e. compressed; with x264 video encoder using constant rate factor (CRF) value of 40 and 50. The HOG features are not always capable of offering a rich set of features when videos are compressed, even shows significant loss of gradients (more significant on fourth image where a higher number of CRF[1] value used to re-encode the video). The usage of compression distorts the edge features which results in assigning almost equal weights to each of the oriented gradients. The loss of edge features results in poor performance of HOG features.

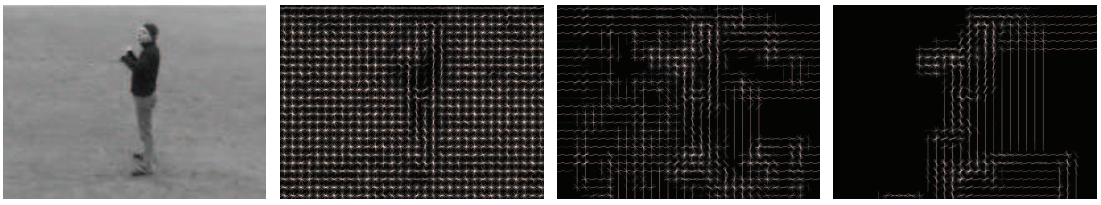


Figure 1.2: An illustration of HOG features (Dalal & Triggs, 2005) with respect to the deterioration of spatial quality. Image sample collected from KTH action dataset (Schüldt et al., 2004).

Beside shape and motion, textural features such as LBP-TOP (Kellokumpu et al., 2008a) and Extended LBP-TOP (Mattivi & Shao, 2009) are also used for recognizing human activities. Their reported performance were promising though the authors

[1]The Constant Rate Factor (CRF) is the default quality setting for the x264 encoder. It keeps up a constant quality by compressing every frame of the same type the same amount, or in other words, its maintaining a constant quantization parameter (QP). The QP defines how much information to "throw away" from a given block of pixels.

admit that they lack explicit motion encoding. Most recently, Kataoka, Aoki, Iwata, and Satoh (2015a) evaluated LBP features for activity recognition and found that they are not effective as shape and motion features for HAR. They also reported that the performance of textures are not so great as shape and motion feature, but they greatly help to improve the performance if they are used in a combined manner. A more detailed review of shape and motion features are given in Chapter 2.

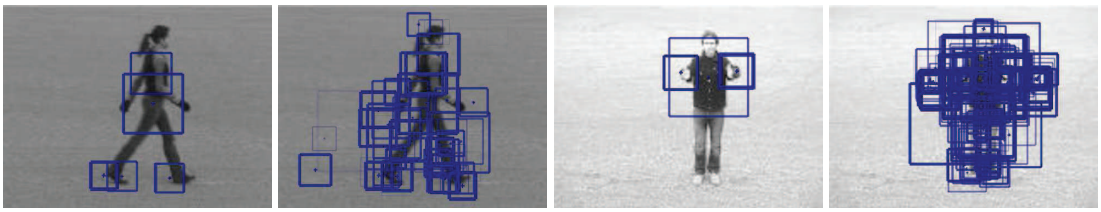


Figure 1.3: An overview of sparse (first and third) and dense (second and fourth) feature selection based on the interest point detection. Figure reproduced from (Willems et al., 2008)

Figure 1.4: Response of feature detectors when videos are downsampled spatially. Sample video frames were collected from KTH actions (Schüldt et al., 2004) and UCF-11 (J. Liu et al., 2009) datasets.

Based on the types of feature selection, current methods can be classified into two type, *sparse* and *dense*. The *sparse* methods select a set of salient features while, *dense* methods select a dense set of features from the action videos, as shown in Figure 1.3. In comparison to their individual performance, dense features perform slightly better than the sparse features, especially when the background is complex (H. Wang

et al., 2009). Both feature selection methods are generally comprises of two main steps: *feature detection* and *feature description*. In *feature detection*, the important features are first detected from videos, and then the visual pattern of detected features are described in *feature description* step. The detection of visual features depends on the image structures which is highly correlated video quality. If the quality becomes low then the selection of important feature regions becomes very challenging. Even fails sometimes, which usually lead towards significant performance drop. Figure 1.4 gives a closer look at the detected features when videos are downsampled spatially. The videos in first row used Harris3D detector (Laptev, 2005) and the second row used improved dense trajectories (H. Wang & Schmid, 2013). The videos in second and third column are respectively downsampled to half and one third resolution of the video showed in column one. The detection of features is greatly affected if the quality of video is compromised.

Owing the drawbacks of existing representation methods, the question comes is - *how do we design a feature representation method that efficiently encodes human activities from low quality video regardless of visual quality?*.

1.3 Research Questions

There are two problems addressed in this thesis. Followings are the research questions arised in this thesis from the lack of framework and feature representation perspectives:

1. How do we incorporate the various features from low quality videos in the activity recognition framework to improve the recognition performance?
2. How do we design a feature representation method that efficiently encodes human activities from low quality video regardless of visual quality?

1.4 Research Objectives

The main goal of this thesis centers on the task of recognizing human activities from low quality videos. Based on research questions showed in Section 1.3, this thesis

aims to achieve the following objectives:

1. To design a feasible framework for human activity recognition under low quality video environment.
2. To develop a new robust representation method for recognizing human activities in low quality videos using spatio-temporal features.

1.5 Scope of Thesis

The scope of research of this thesis is to recognize human activities from low quality videos using spatio-temporal features. There is no formal definition of low video quality in literature, but usually the videos that are affected by the factors or problems such as low spatial and temporal resolution, noise, motion blurring, camera ego-motion and compression artifacts are considered "low quality". However, as a scope of this thesis, only low quality videos that are affected by "low spatial resolution", "low sampling rate", "compression artifacts" and "motion blur" are considered. The type of human activities that this research want to recognize include: (1) single person activities and (2) person-object interactions. The potential application domain for human activity recognition from low quality videos includes but not limited to web, aerial, crowd scene and sports videos.

1.6 Contributions of this Thesis

The contributions of this thesis are described as follows:

1. A joint feature utilization based framework or pipeline for recognizing human activities is proposed which combines various spatio-temporal features representations inside its components. While existing frameworks use only shape and motion features, proposed framework also uses textural features to describe the human activities, which is a novel proposition that increase the recognition performance in low quality videos by a good margin.
2. A spatio-temporal mid level feature bank (STEM) for low quality video is proposed. The feature bank comprises of a trio of local spatio-temporal feature that

encodes shape, motion and textures from low quality videos. Textural features are proposed to extract discriminately from 3D salient patches.

1.7 Preview of the Chapters

The rest of thesis is organized as follows: Chapter 2 discusses the recent methods developed for human activity recognition. A comprehensive taxonomical review on spatio-temporal features is presented to discuss various category of methods proposed in recent literature. A further discussion on the methods specifically designed for low quality video is also presented in this chapter. Chapter 3 discusses the datasets and their evaluation criteria's. Since there is a lack of dataset for low quality videos so, various methodologies developed for the creation of low quality videos. Chapter 4 discusses the joint spatio-temporal feature utilization framework and its several components for activity recognition in low quality videos. Chapter 5 discusses the proposed spatio-temporal mid level feature bank (STEM) and its various parts for recognizing activities in low quality videos. Finally, Chapter 6 will discuss the concluding remarks, limitations and future work directions.

CHAPTER 2

LITERATURE REVIEW

There is a substantial research on vision based human activity recognition reported in literature, with most of them are designed to process only good quality videos. A number of surveys (Aggarwal & Ryoo, 2011; Ke et al., 2013) are also available, but there is no particular survey that reviews the area of human activity recognition in low quality videos. This is due to the lesser availability of works (mostly are on low resolution only) on low quality videos to warrant a survey.

Most early efforts in human activity recognition have used state-space and template matching models. Aggarwal and Cai (1997) and L. Wang, Hu, and Tan (2003) thoroughly reviewed various state-space and template matching model based methods with their limitations. Efforts in reviewing earlier methods from different perspectives are also available in literature such as level of hierarchy (A. F. Bobick, 1997), feature space dimensionality, i.e. 2D and 3D (Gavrila, 1999), body pose, body parts, action grammars and primitives (Moeslund et al., 2006), complexity level of action and activities (Turaga et al., 2008), and local and holistic nature of feature estimation (Poppe, 2010).

A recent survey by Aggarwal and Ryoo (2011) comprehensively summarized and compared the most significant progress in this field of human activity recognition. Based on whether the action is recognized from input image directly, they divided existing works into two major categories: single-layered and hierarchical approaches. Both are further sub-categorized depending on the feature representation and learning methods, as shown in Figure 2.1. Two additional surveys by Ke et al. (2013) and S. Vishwakarma and Agrawal (2013) further extended the work of Aggarwal and Ryoo (2011) by adding few works on object tracking and activity classification. Ke et al. (2013) also reported the existing methods based on their level of representation i.e

high, mid and low. X. Xu et al. (2013) gave a brief review about recent template matching and state-space approaches. Recent surveys mostly focused on specific type of methods rather than reporting comprehensively across the board. As an example, Dawn and Shaikh (2015) discussed spatio-temporal interest point based methods while H. Xu et al. (2015) gave a brief overview about dense trajectory based methods.

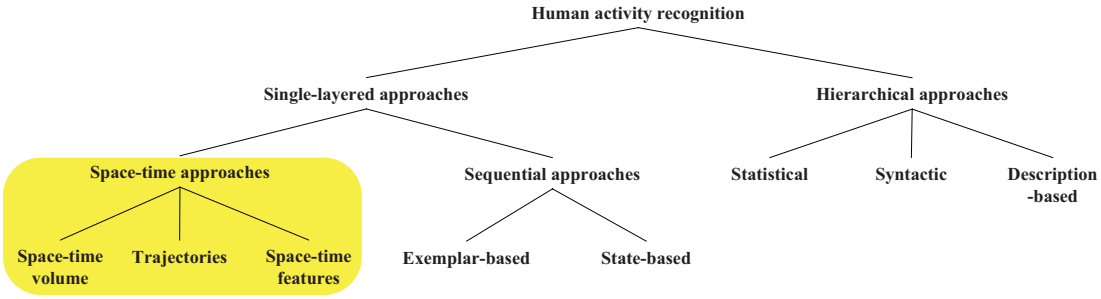


Figure 2.1: Hierarchical approach based taxonomy of human activity recognition methods. Figure reproduced from Aggarwal and Ryoo (2011).

In the aspect of features generation, there are mainly two distinct category of works available in literature namely, *handcrafted* features and *unsupervisedly learned* features. The majority of the methods reported in the literature is hand-crafted while unsupervisedly learned methods have recently gained popularity due to their good results in dealing with large-scale visual data. In order to report the progress of hand-crafted methods, similar taxonomy as in Aggarwal and Ryoo (2011) is used, but in line with the research scope of this thesis mentioned in Section 1.5, only space-time approaches are discussed. On the other hand, from the taxonomical point of view, unsupervisedly learned space-time methods can be divided into four categories, namely (i) Spatio-temporal networks, (ii) Multi-stream networks, (iii) Deep generative networks and (iv) Temporal coherency network. Since all unsupervisedly learned features are operating in a similar domain (i.e., spatio-temporal) as space-time features so, they will be described together.

In this chapter, space-time approaches are first introduced in Section 2.1, then progress in each of its categories namely, space-time volumes, space-time trajectories and space-time features are discussed. Section 2.2 discusses some existing works that

are related to low quality video and Section 2.3 summarizes this literature review with important observations that provides the key directions of this research.

2.1 Space-time approaches

Video consists of a temporal (T) sequence of 2-D spatial (XY) images, or equivalently a set of pixels in 3-D XYT space. Therefore, a sequence of image frames, i.e. video; can be considered as a space-time or spatial-temporal volume, and this volume contain necessary information for human beings and machines to recognize the actions and activities in the volume. Based on this assumption, various representation and correspondence matching algorithms have been put forward to compactly characterize the underlying motion patterns.

In this section, the progress of various categories of space-time approaches is discussed. Except for the methods that use raw volume as a feature such as A. F. Bobick and Davis (2001), all categories discussed uses motion-related information to characterize human activities.

2.1.1 Space-time volume-based approaches

Space-time volume-based (STV) approaches are one of the most popular and earliest methods for activity characterization in the video. STV methods, model human activities directly from the 3D volume (XYT) and measure the similarity between various volumes. Many methods have been proposed in literature for finding accurate volume similarities. The most intuitive methods would use the entire 3-D volume as feature or template, and match unknown action videos to existing ones. However, since activity modeling directly from raw video frame values suffers from noise and meaningless background information, and therefore, some effort has been made to model the foreground movements. Based on the feature generation strategies, existing STV approaches can be grouped into three groups: (i) Template based, (ii) Silhouette and skeleton based, and (iii) others. The details of each group and related notable methods are briefly described below:

Template based methods

Generally, template based methods rely on the template (created from video) similarity matching between different video samples. If two templates are close or similar with each other they considered belonging as to the same action class. A. F. Bobick and Davis (2001) first proposed the idea of a template based action modeling from an STV (XYT). They proposed 2-D (XY) Motion history image (MHI) and Motion energy image (MEI) from the 3-D image stack (i.e., video) for modeling foreground actions. They managed to recognize simple human activities such as sitting and waving in real-time by estimating the similarities between the MHI and the MEI template's images. However, modeling of actions from complex scenes using their method is difficult, since it is very hard to differentiate between foregrounds and backgrounds. The sample images of MHI and MEI are shown in Figure 2.2.

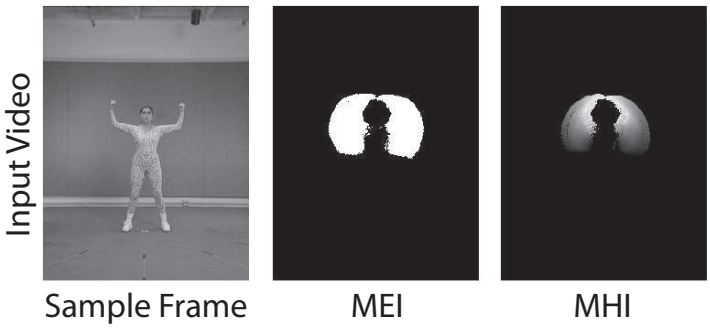


Figure 2.2: Sample illustration of Motion history image (MHI) and Motion energy image (MEI) (A. F. Bobick & Davis, 2001). Figure reproduced from A. F. Bobick and Davis (2001).

Inspired by the success of MHI and MEI in characterizing actions, various approaches have tried to extend it for complex action recognition by incorporating appearance information. For example, Babu and Ramakrishnan (2004) used coarse MHI and motion flow history image (MFI), while Meng et al. (2006) used multi-valued differential image (MMHI) (Ogata et al., 2006) and Motion Gradient orientation (MGO) (Bradski & Davis, 2002) with principal component analysis (PCA), Hu et al. (2009) used foreground image and histogram of oriented gradients (HOG), while Qian et al.

(2010) used MEI based contour coding and object features, and Roh et al. (2010) used volume motion template (VMT) and followed by Murakami et al. (2010) who used directional MHI (dMHI). All of these methods detect additional shape or appearance features, and combined them with MHI or MEI templates in order to increase their robustness. However, these methods may not be suitable if shape information is distorted by noise or other factors.

There are a few works available that uses motion information to further improve the MHI and MEI templates. For example, M. A. R. Ahad et al. (2011) used dMHI and MEI templates for description of SURF (Willems et al., 2008) features. dMHI calculates optical flow feature from four directions – top, bottom, left and right. Shao et al. (2012) used Pyramid Correlogram of Oriented Gradients (PCOG) features from localized action parts based MHIs and MEIs. J. X. Cai et al. (2013) used region of interest (ROI) based Pixel change history (PCH) features for removal of background motions. Dogan et al. (2015) used 3D volume motion templates (VMT) calculated from video *tracklets*. They also did some pre-processing for noise removal and used HOG3D feature (Klaser et al., 2008) to describe VMT. Tsai et al. (2015) used optical flow motion history image (OF-MHI) features. While MHI only represents static motion information, the OF-MHI encodes more dynamic motion information, even if they are sensitive to the noise problems. The usage of motion information on MHI and MEI greatly improved their performance, but in contrast with realistic videos where various complex problems such as high camera motion, jitter and blur are addressed, these methods may not be a suitable choice.

Few works also used textural features to describe MHI and MEI templates, such as Kellokumpu et al. (2008b) which used local binary patterns (LBP) (Ojala et al., 2002) for description of MHIs and MEIs. Ahsan et al. (2014) also used LBP images, but they used it to describe dMHI templates. Their combined use of LBP and dMHI templates helps to achieve the robustness across general purpose activity recognition problems. However, LBP features do not work well if video quality is high. Also, in these methods, LBP was used only to extract motion structures which is not useful, if

the videos are suffering from any explicit motion problems.

There are few recent works available inspired from the MEI methods. For example, L. Wang et al. (2013) proposed spatio-temporal orientation energy (HOE) for characterizing human activities. For estimation of HOE, at first a third derivate of 3D Gaussian was first applied unto of each image pixel value. They used eight different normalized energy image for estimating robust orientation map. The energy images representative of motion saliency maps in various directions. However, it does not perform well in situations where video is complex, such as in the HMDB51 dataset (Kuehne et al., 2011).

There are also some methods available that adopts the template based philosophies for feature representation, such as Laplacian Pyramid Coding (Shao et al., 2014) and genetic programming driven STV features (L. Liu et al., 2016) etc.

Silhouette and body parts based methods

Silhouette based methods rely on the shape extraction from STV. Most of the methods in this family are inspired from template based methods, i.e. MHI and MEI; and tries to solve the problems faced by those methods by only considering the silhouette information. One of the early work by Han and Bhanu (2006), addressed the problems of MHI and MEI in modeling cyclic and self-occluded motion, and proposed gate energy image (GEI) to encode walking type of videos. The sequences of binary silhouettes is first extracted from the 3D volume and normalized to get fixed shape sizes across the image frames. This is done by averaging the binarized sequences of silhouettes. This normalized representation is considered as a GEI. A method for learning distortion free GEI from various activities was also proposed. The similarity between various GEIs is measured by finding the minimum distance between the test and training samples.

Another notable early work is Action Energy Image (AMI) (Chandrashekhar

& Venkatesh, 2006) where, silhouettes are extracted using Gaussian mixture model density and background subtraction. Then a 1D Fourier filter is applied on the 3D volume to obtain the 2D AEI. The PCA is used to reduce the dimensionality of feature in the feature space. The Euclidean distance metric was used to find the similarities between the test and training samples.

Motivated by GEI, many researchers proposed methods to improve the silhouette information. For example, W. Kim et al. (2010) proposed accumulated motion image (AMI) to represent spatio-temporal features of occurring actions. Folgado et al. (2011) subdivided 2D silhouettes into 6 distinct equal regions into block for feature extraction. Aminian Modarres and Soryani (2013) proposed the idea of body posture graph (BPG), where body parts are extracted from video using EBF kernel (an extension of RBF kernel). Gupta et al. (2013) used rule based MHIs (R-MHI, G-MHI and B-MHI) for silhouette extraction in RGB color space. D. Vishwakarma and Kapoor (2015) used shape and rotation features from silhouette. Chaaraoui et al. (2013) used silhouette contour points based key-pose features. All of these methods have made efforts in efficient feature extraction from silhouettes. However, these methods are only applicable to the videos that are captured in a controlled environment, and have simple background. Also, identifying the differences between speed of actions, such as *walking* and *slow running*, are a bit difficult with these kind of methods.

Beside conventional efforts in modeling activities from silhouettes, researchers have also focused on different directions, such as semilatent topic models (STM) (Y. Wang & Mori, 2009), independent component analysis (ICA) and over-complete ICA (S. Zhang et al., 2014).

Silhouette based methods suffered from many problems, with one of the big problems being complexity in modeling of complex activities such as *dancing*, *soccer playing* and *jogging* from silhouettes are challenging. *Body-part based methods* minimizes these problems with sampling features from only selected silhouette regions. They use discriminative parts of the body silhouette region or various locations of

STV for modeling activities. Among popular methods in this family, histogram of oriented rectangles (Ikizler & Duygulu, 2009), cumulative skeletonized image (Ziaeefard & Ebrahimnezhad, 2010) and Histogram of body parts (C. Wang et al., 2013) are included.

Histogram of oriented rectangles (HOR) represents local activity dynamics in an activity sequence with oriented rectangular patches extracted from human silhouettes. Cumulative skeletonized image (CSI) represents frame-specific motion information across the time. The CSI is usually expressed as a 2-D angular or distance histogram. Histogram of body parts (HBP) is an extension of Yang and Ramanan (2011). At the beginning, body parts are estimated from STV, and divided them into five groups. Then, spatial and temporal data mining is applied to find the co-occurrence between body parts in spatial and temporal domain. The distinctive body parts are then represented as a histogram of body parts. However, all of these methods may be applicable to only controlled environment videos such as KTH. The successful extraction of silhouettes from complex backgrounds is very challenging.

Other methods

Beside template, silhouette and body part based methods, efforts in other directions have also occurred in literature. For example, T.-K. Kim and Cipolla (2009) extended Canonical Correlation Analysis (CCA) to measure video-to-video similarities. The method acted upon video volumes avoiding the difficult problems of explicit motion estimation, and provided a way of spatiotemporal matching that is robust to intraclass variations of action due to CCA. C. Liu and Yuen (2010) use principal component analysis (PCA) to a salient action unit (SAU) (i.e., one cycle of repetitive action in a video), and AdaBoost classifier was used to classify the action in a query video. Cao et al. (2009) provided a new way to combine different features using a heterogeneous feature machine (HFM).

Seo and Milanfar (2011) used 3D LSK (Shechtman & Irani, 2007) for extrac-

tion of features from densely extracted patches from 3D volume for one-shot learning. B. Li et al. (2011) proposed dynamic sub-space angles (DSA) where, temporal information are first represented using Hankel matrix and then canonical correlations among the subspaces around columns of Hankel matrices are estimated for forming final features. Harandi et al. (2013) used Grassmann manifold for encoding activities from complex videos. Fu et al. (2013) used Principal Geodesic Analysis (PGA) on Grassmann manifold for reduction of computational overhead. Fu et al. (2013) used iterative tensor decomposition method for classify tensor video representation in an unequal length of time. Zhou and De la Torre (2016) used Generalized Canonical Time Warping for alignment of activities of different peoples in video clips.

2.1.2 Space-time trajectory-based approaches

Trajectories are usually constructed by tracking body joint points or interest points across the STV. Trajectory based methods, believes that the observation of these positions is enough for estimation of body motion (Johansson, 1975). Various representations and algorithms were proposed in literature to match the trajectories for action recognition. Based on the type trajectory generation, existing methods can be categorized into three types: (1) Salient Trajectories, (2) Dense Trajectories, and (3) Other methods. We discuss notable methods from each of the groups, in the following discussion.

Salient trajectory based methods

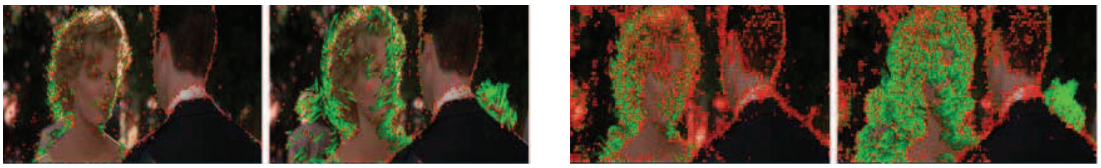
Salient trajectory based methods represents human activities by tracking feature points over time. Generally, at first, they detect salient points from the image frames and track them across the consecutive frames. The features such as shape and motion aligned with different trajectories are represented using respective feature descriptors. A typical illustration of salient trajectory based method (Messing et al., 2009) is shown in Figure 2.3. According to literature, the earliest methods were mostly focused on trajectory design or generation, while the later and recent methods deal with improving the existing methods by removal of irrelevant trajectories.

Figure 2.3: Salient point trajectories using feature points tracking by KLT tracker (Lin et al., 2009) across consecutive image frames. Figure reproduced from Messing et al. (2009).

One of the earliest notable work is by Messing et al. (2009). They extracted feature trajectories by tracking Harris3D (Laptev & Lindeberg, 2003) interest points using a KLT tracker (Lin et al., 2009), and represented them as sequences of log-polar quantized velocities. Inspired by their work, Matikainen et al. (2009) also used the KLT tracker for tracking feature points, but they followed a different approach for feature representation. They created trajectory snippets for representation of activity portions from video samples. However, Harris feature points and KLT tracker are limited to scale changes. Sun et al. (2009) addressed this problem and proposed to use SIFT detector (Lowe, 2004) for detection feature points, and pairwise SIFT matching for tracking of SIFT points between consecutive frames. For the description of tracked feature points, SIFT features were extracted from each frame and then averaged.

The ideas of simple point trajectories is not sufficient to deal with situations where large motion structure is involved. For example, SIFT trajectories is not sufficient to represent large motion trajectories. There are few methods which addressed this problem, e.g. Sun et al. (2010) extended the idea of SIFT trajectories to the large trajectory generation problem by incorporating the idea of combing SIFT points, KLT tracked points, triangular mesh and random feature points. Their approach was able to perform well in situations where shape deformation, occlusion and camera motion is involved. Raptis and Soatto (2010) used common feature point detectors such as

Harris and SIFT to detect features, and tracked them using a contrast-based translation sensitive feature tracker for obtaining the trajectories, called 'tracklet'. The time series of histogram (Ho) and average (Ao) of gradients (G) and optical flow (F) were used to describe the tracklets. Rubinstein et al. (2012) used LDOF tracker (Sundaram et al., 2010) for generation of tracklets. They combined short-term length trajectories for generating long-term trajectories by applying a divide and conquer approach. A block-based feature descriptor is used for describing trajectory aligned features.



(a) SIFT keypoints trajectory (b) Dense trajectory

Figure 2.4: Example SIFT and dense trajectory generation from consecutive video frames. Figure reproduced from H. Wang and Yi (2015).

If video scenes become complex and is affected from noise or camera motion, sometimes some irrelevant trajectories are also generated from video that manipulates the discriminative capacity of trajectories, hence, may lead to poor performance. So some efforts are given in pruning the irrelevant trajectories from videos. Yi and Lin (2013) used point trajectories based on image saliency map. J. T. Zhang et al. (2014) used SIFT features and tracked them using the SIFT flow (similar like optical flow). Their generation of trajectories is similar to the Dense trajectories (H. Wang et al., 2011). Inspired by Improved Dense trajectories (IDT) (H. Wang & Schmid, 2013), H. Wang and Yi (2015) proposed SIFT trajectories which also use dense warped flow field for the generation of trajectories. Similar to IDT, estimation of warped flow allows SIFT trajectories to track more robust trajectories free from irregular motion.An illustration of comparison between SIFT and dense trajectories is shown in Figure 2.4.

Dense trajectory based methods

H. Wang et al. (2011) first introduced the dense trajectories (DT) for activ-

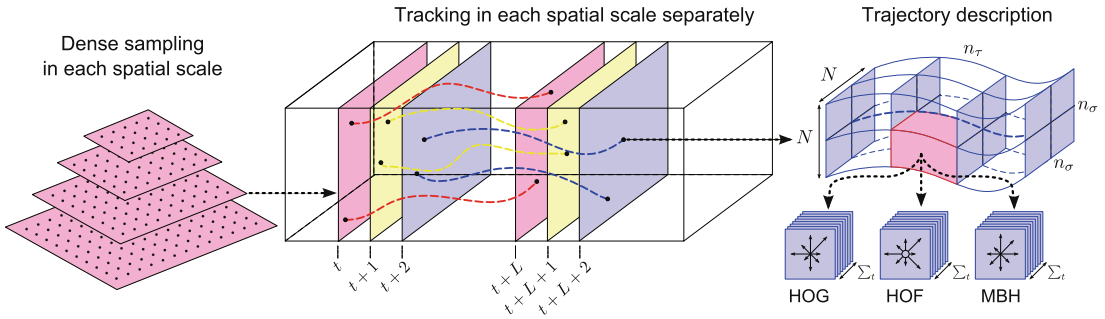


Figure 2.5: An overview of dense trajectories where feature points are tracked across the video frames and then described by various feature descriptors. Figure reproduced from H. Wang et al. (2011).

ity recognition. They sampled dense points from each frame and tracked them based on displacement information from a dense optical flow field. Trajectory aligned descriptors such HOG, HOF and MBH (motion boundary histogram) are computed for activity modeling. An illustration of the DT is shown in Figure 2.5. DT showed good results across many complex video datasets such as KTH (94.2%) and UCF-YouTube (84.2%). However, feature tracking from dense optical flow field also considers camera motion that may affect in discriminative feature sets. Addressing this issue, they further extended their work of dense trajectories to improved dense trajectories (IDT) (H. Wang & Schmid, 2013). They used warped flow for estimating the irrelevant trajectories. The estimation of warped flow helped feature descriptors to improve their performance. IDT obtained good performance across many challenging datasets such 57.2% for HMDB51 and 91.2% for UCF50 (Reddy & Shah, 2013).

However, dense trajectories do not consider any relationships between foreground and background which may be a problem if action videos with complex scenes is involved. So, some efforts are given to establish this relationship. For example, Jiang et al. (2012) proposed to use local and global feature points for modeling motion relationship between the moving object and the respective background. This also allows trajectories to be free from irrelevant camera motions. Gaidon et al. (2012) proposed cluster tree representation of dense trajectories to hierarchically segment motion parts for modeling activities in complex videos. However, this method will be only feasible if videos are temporally long such as KTH (Schüldt et al., 2004). For short-term videos

with simple background, it is quite difficult to have a large number of trajectories to produce a cluster-tree.

Though many suggested that the removal of camera motion removes unnecessary trajectories, some methods available in literature relies on explicit trajectory selection through saliency maps or temporal similarities. Recent works are mostly falling into this category. For example, M. Jain et al. (2013) extracts trajectory aligned dominant motion information and describe them with DCS feature descriptors. Instead of estimating homographies for removal of camera motion as in IDT, they focused on compensating the dominant motion starting from the tracking stage to the description stage. Peng et al. (2013) used optical flow based motion boundary image for pruning irrelevant DTs. O. Murthy and Goecke (2013) proposed ordered trajectories by removing irrelevant DT points based on similarity (using distance metric) between feature points in consecutive frames. In O. R. Murthy and Goecke (2015), they further evaluated their idea on improved trajectories (H. Wang & Schmid, 2013) and demonstrated higher performance across various datasets, mostly on large-scale complex datasets including HMDB51 (58.8%) and UCF50 (92.1%). The removal of irrelevant trajectories strengthens the discriminative capacity of the features, and also at the same time it reduces a lot of the computational overhead.

Very recent works use trajectory features to improve their performance. Among notable methods, L. Wang et al. (2015) used IDT features for pooling deep convolutional spatio-temporal feature maps. They separately extract dense trajectories and CNN feature maps, and calculate features from feature maps by pooling (sum) over trajectories (TDD). This strategy (TDD and TDD+iDT) achieved competitive performance among HMDB51 (65.9%) and UCF101 (Soomro et al., 2012) (91.5%) datasets. Ma, Bargal, et al. (2015) used web images for fine tuning CNN features and combined them with improved dense trajectory (iDT) based features for achieving robustness towards motion information. The trajectory feature ignores the object or shape relationships in the frames and highly focuses on detection of trajectories based on motion information. In that case, CNN feature maps encode rich shape features which

can be used for trajectories to improve their performance. This method makes use of both CNN and trajectory aligned features such as HOG, HOF and MBH managed to achieve higher performance in uncontrolled complex web videos i.e., 91.1% for UCF101 dataset.

Other methods

Beside salient point and dense trajectories, efforts in other directions for the detection of feature trajectories are also available in literature. Ali et al. (2007) proposed trajectory features based on chaotic theory for modeling non-linear activity dynamics. Bregonzio et al. (2010) used ROI based SIFT feature point trajectories where, frame-wise global motion estimation and collaborative feature selection are used. Raptis et al. (2012) used clustered trajectories to determine activity part instances of a particular class which are later grouped together using latent variables for finding the similar parts. Cho et al. (2014) used group sparsity based trajectory clustering for selection of local key-trajectories. Ramana Murthy et al. (2014) used mid-level body part based trajectories for modeling activities. Among all these, dense body part trajectories managed to achieve a competitive performance across a number of complex datasets such as HMDB51 (59.4%) and UCF50 (92.1%).

2.1.3 Space-time features based approaches

Generally, space-time feature (STF) based approaches model activities based on space-time features and recognize activities by considering the match between those features. STF samples features from STV. Based on literature and feature generation, STF feature are grouped into three: (i) Interest point based features (ii) Densely sampled features, and (iii) other features. Below we discuss some notable methods from each group.

Interest point based features

The application of local interest point features in action recognition is extended

from object recognition in images (Cheng et al., 2015). The local features refer to the description of points and their surroundings in the 3-D volumetric data with unique discriminative characteristics. In terms of the density of extracted feature points, the representation of local feature approaches can be divided into two broad categories: sparse and dense. The Harris3D detector (Laptev, 2005) and the Dollár detector (Dollár et al., 2005) are representative of the former, and optical flow-based methods the latter. Most algorithms in this family are derived from them.

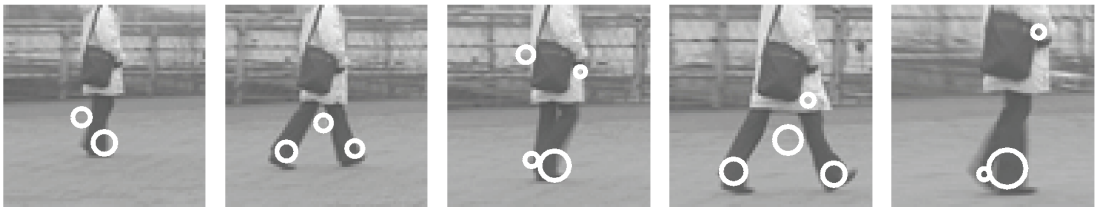


Figure 2.6: Detection of Space-time interest points (STIPs) in 'Walking' video. Figure reproduced from Laptev (2005).

Laptev and Lindeberg (2003) proposed the idea of space-time interest point (STIP). They extended the Harris detector (Harris & Stephens, 1988) used popularly in image domain to space-time, namely the Harris3D detector in order to detect salient points from videos. To describe visual pattern across the STIPs, they used normalized spatio-temporal Gaussian derivatives varying scales at which STIP was detected. In Schüldt et al. (2004), they further evaluated the idea of STIP using the KTH dataset and achieved good performance on KTH ($\approx 75\%$). They used multiple feature descriptors such as such as histogram of local features (HistLF), and histogram of normalized spatio-temporal gradients (HistSTG) in order to describe visual patterns across the interest point (IP). However, the feature descriptors they used are not efficient at encoding local motion and shape features, especially if videos with complex environments are considered. The detection of STIPs using Harris3D in video frames is shown in Figure 2.6.

Laptev et al. (2008) utilized STIPs to recognize activities from movie clips. In order to cope with video complexities, they introduced HOG and HOF feature de-

scriptors to describe IPs. HOG descriptor describes structural patterns, while HOF describes the motion patterns around the STIPs. They managed to retain competitive result for KTH dataset (91.8%). Klaser et al. (2008) also used STIP but extended the idea of HOG to HOG3D where, the gradient orientations are quantized in regular polyhedrons. The idea of HOG3D showed good performance across few popular activity datasets including KTH dataset (92.6%). A recent evaluation by (H. Wang et al., 2009) further evaluates the performance of STIP on various feature descriptors, and demonstrated their effectiveness on various complex datasets such as UCF sports (Rodriguez et al., 2008) and Hollywood2 (Marszalek et al., 2009).

Inspired by space-time interest points (Laptev & Lindeberg, 2003), more researchers also proposed their idea of detecting STIPs (most of the ideas are just an extension of the Laptev's idea). The STIP methods only selects the salient points from video, but it does not always produce enough features to describe the activity especially if videos with complex background are considered. The subsequent works have tried to resolve this issue by introducing the concept of dense STIP such as Dollár et al. (2005) and Willems et al. (2008).

Dollár et al. (2005) proposed the idea of using the spatio-temporal cuboid detected directly from STV. They used temporal Gabor filters to detect STIPs across the video and, extracted the cuboid around it. On the other hand, Willems et al. (2008) proposed the Hessian detector which is an extension of Hessian saliency measure (Lindeberg, 1998) used for detection of image blob. To describe the detected STIP they used the ESURF (extended SURF) descriptor. However, according to the evaluations by H. Wang et al. (2009), cuboid and Hessian detectors does not perform well in complex datasets such as UCF Sports and Hollywood2 as compared to the HOGHOF descriptors.

In complex videos, sometimes many unnecessary background STIP are also detected which often reduce the discriminative capacity of the activities. To cope with this issue, they proposed to improve the STIPs by removing irrelevant STIP features

detected in the background. Among popular methods, J. Liu et al. (2009) used statistics of motion to get more stable motion and clean shape features detected by the Dollár detector. They used a PageRank algorithm for selection of shape features. Their method demonstrated $\approx 8\%$ improvement of performance on the YouTube dataset (J. Liu et al., 2009), in comparison with the baseline STIP features. Chakraborty et al. (2012) proposed background suppressed STIPs. Their approach achieved a competitive performance, i.e. 96.35% for KTH, and 86.98% for YouTube. Gilbert et al. (2011) used mined STIPs. Q. Wu et al. (2013) used saliency map (Harel et al., 2006) to prune STIPs. However, the use of saliency map may not produce effective results on complex datasets, for example, for the YouTube dataset it achieves 83.1% while other methods performed better than this.

While most methods considers only shape and motion features, D. Zhao et al. (2013) proposed to use appearance features for description of STIPs. They used a optimized 3D shape context descriptor for description of appearance. However, their method only performs well for controlled videos such as KTH.

Another notable method is Color STIP by Everts et al. (2014). They incorporated chromatic information into the detection of STIPs. Popular STIP detection methods such as Harris3D and Cuboid mainly depended on intensity information and may fail if any disturbing motion is happening in the frame. Incorporating detected features from multiple color channels will solve this issue by allowing feature detection from various color channel perspective. A figure of STIP detectors after incorporating chromatic information is shown in Figure 2.7. The use of color interest point helps to increase the performance in many complex video samples. It manage to detect more feature points in comparison with intensity based detectors. However, if spatial resolution is distorted too much, such as during high compression, and gray scale video is used, then color detectors may not be a suitable choice.

Instead of concentrating only on feature representation, a few researchers also have given the efforts in feature encoding. One notable work is by Peng et al. (2014).

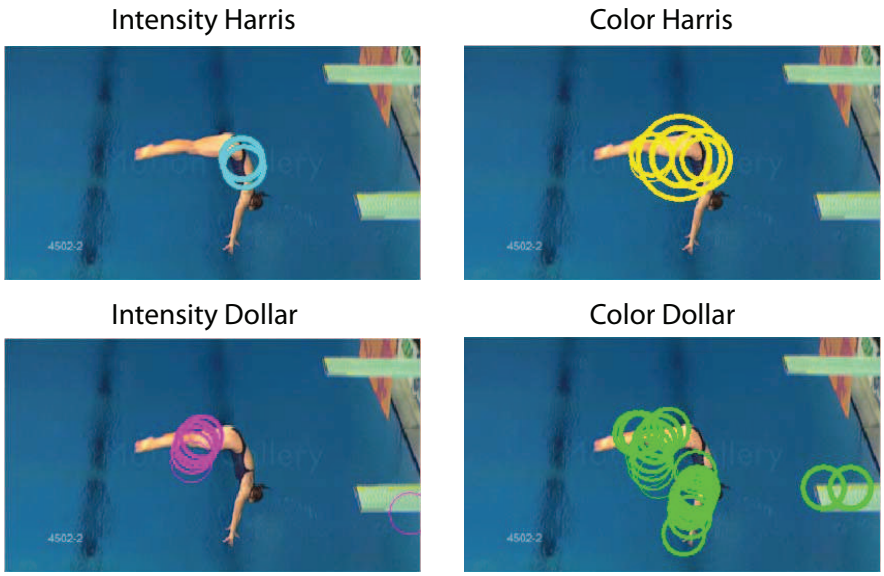


Figure 2.7: Comparison between intensity and color based Harris3D (Laptev & Lindeberg, 2003) and Dollar (Dollár et al., 2005) detectors. Figure reproduced from Everts et al. (2014).

They evaluated STIP features on large-scale datasets using various encoding methods, and demonstrated various feature fusion rules. In their evaluations, across many available encoding and fusion methods, fisher vector (FV) (a super-vector based method) performed consistently better across various datasets but it also will lead towards a higher computational cost. They have also demonstrated that the use of the proper normalization method will further help to improve the performance.

STIP generated features are lack of object properties, they only capture certain visual patterns from an object and do not care sometimes if it not even the object. One recent work by O. Murthy and Goecke (2015) addressed this issue, and used deep convolutional neural network (CNN) based object features for improvement of the performance of STIP based feature. They make use of late CNN layer features, i.e. FCs and softmax; to supplement HOG features in order to improve the discriminativeness of the structural features. Their method achieved competitive performances on HMDB51 (50.5%) and UCF101 (84.5%).

Dense feature sampling based methods

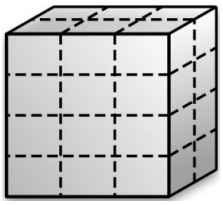


Figure 2.8: Regular blocks in a video volume is defined and used for the extraction feature descriptors.

Despite of the availability of selective feature selection methods, some authors also proposed te extraction of dense features, such as H. Wang et al. (2009). They proposed to extract multi-scale video blocks at regular positions (in both spatial and temporal domain). Since the spatial domain has more importance (temporal information estimation is mainly based on spatial information contained in each image frame) than the temporal domain, they extracted features with 8 different spatial scales and only 2 temporal scales (in total sampled a video 16 times considering all scales). Each block with varying scales are then described using various feature descriptors such as HOG, HOF and HOG3D. The method showed comparatively good performance for complex videos. However, dense sampling requires high computational resources if large-scale evaluations are required to be undertaken and also, in case of controlled environment videos such as KTH it may not be a good choice (H. Wang et al., 2009). A sample of video blocks at regular positions is shown in Figure 2.8.

Since dense sampling also considers many irrelevant features from the video so, few efforts have been made to prune them. For example, Vig et al. (2012) used saliency maps. Their method was able to improve the performance of the baseline method (H. Wang et al., 2009), and obtained a mean average precision (mAP) of 54.5% for Hollywood2 dataset. M. Chen et al. (2015) proposed bounding-box appended HOG3D (Klaser et al., 2008) patches for removal of irrelevant background features. Nguyen et al. (2015) extracted multiple saliency maps and combined them into one map for prediction of the final saliency map. The final feature maps are then used to pool the dense video features in the spatio-temporal direction with the concern of visual attention. They managed to get good performance across a number of datasets such as YouTube (87.9%). However, generation of many saliency maps for

single map prediction is computationally very expensive.

The above mentioned methods make use of straightforward regular grid based densely sampled feature pruning using feature maps or human-centered bounding boxes, but there are also methods available that randomly extracts dense blocks from video such as local part model (Shi et al., 2013) and gradient boundary histogram (Shi et al., 2015). Local part model (LPM) extracts overlapped multi-scale dense spatio-temporal patches from the video that concerns with two main factors: root and associated parts. The root patches are extracted from half-resolution video while parts are extracted from full-resolution video at the reference point of root. The patches are then described using HOG3D descriptors. Shi et al. (2015) also made use of LPM method, but they used gradient boundary histogram (GBH) – a new and simple gradient based descriptor for feature description. However, computation parts and roots are computationally expensive to handle so, in Shi et al. (2016) they further improved it by considering these two patches separately.

The dense sampling methods are computationally very expensive to handle and mostly not suitable for real-time. So, some efforts have given also to optimize them for real-time recognition. For example, Uijlings et al. (2014) used HARR features instead of gradients for shape descriptor estimation and classic Horn-Schunk optical-flow for motion descriptor estimation from regular densely sampled grids and further reduce them using PCA. They also presented some empirical evaluations on densely extracted features for real-time use.

Unsupervisedly learned methods

While handcrafted features are carefully hand engineered to obtain the efficient features, Unsupervised learned methods learn features directly from visual data. There are many unsupervised learned methods for activity recognition is available in literature, from which, deep learned methods such as Convolutional Neural Networks (CNN) (LeCun et al., 1998) have become popular in recent years. Generally, dealing

with these methods is computationally very expensive, and efficient feature learning with these methods requires large-scale data samples. There are many deep learning architecture available in literature, from taxonomical point of view, they can be classified into four categories, namely (i) Spatio-temporal networks, (ii) Multi-stream networks, (iii) Deep generative networks and (iv) Temporal coherency networks (Herath et al., 2016). Below we discuss some notable methods form each type.

Spatio-temporal networks treat videos as an STV. A direct approach by (Ji et al., 2013) uses CNN architecture for convolution with temporal information. It applies convolution directly on spatial and temporal domain of video, hence encodes motion information, i.e. with the help of optical flow images. It outperforms traditional frame-based CNNs. However, it requires fixed size STV as an input which may result in loss of spatio-temporal information. There are few works that address the CNN temporal resizing problem. For example, Yue-Hei Ng et al. (2015) suggests temporal max-pooling and Karpathy et al. (2014) used the concept of 'slow fusion' for temporal awareness in CNN network. Some researchers also use recurrent networks for utilization of temporal information. For example, Donahue et al. (2015) used Long-term Recurrent Convolutional (LRCN) network to capture the temporal evolution for video captioning, Baccouche et al. (2011) use 3D CNN learned features on a Long-short Term Memory (LSTM) network for activity classification.

Multi-stream networks treats appearance and motion information separately for activity modeling from STV (Simonyan & Zisserman, 2014). There are many notable works in this family. Simonyan and Zisserman (2014) is one the first who introduced the idea of multi-stream network. They separate their network into two streams, namely spatial and temporal stream. For each stream they have used VGG-16 deep architecture (Chatfield et al., n.d.) where, spatial stream accepts raw video frames and temporal stream uses optical flow images. They achieved a good performance in two popular large-scale datasets, namely UCF-101 (88.0%) and HMDB51 (59.4%). Inspired by the performance of two stream network, L. Wang et al. (2015) used it for representation of improved trajectory features. Z. Wu et al. (2015) extended

to three stream network by adding a third stream using audio signal into the network. However, multi-stream networks requires to train multiple network, which is computationally very expensive to handle.

Deep generative networks are used for deducing the underlying data distributions. These methods are from the family of 'Auto-encoders'. These network models efficiently learn temporal sequences. One of the recent methods by Yan et al. (2014) introduced 'Dynencoder' for synthesizing dynamic textures in STV. It was successful in encoding the temporal evolution of video. Another recent work by Srivastava et al. (2015) used LSTM auto-encoder for encoding long-term activity cues. It used two RNNs (recurrent neural network), one as an encoder and another as a decoder. It also can be used for early prediction in video. However, these networks are very expensive to train, requires a lot of computational resources. There are some deep generative networks that use adversarial philosophy for video modeling. One notable work is by Mathieu et al. (2015) who uses an adversarial methodology for the removal of CNN pooling layers during the training of multi-scale CNN.

Temporal coherency networks are used to learn features from unlabeled videos. A recent work by Misra et al. (2016) used temporal coherency for estimating the body posture for activity recognition. Particularly, a 'Siamese Triplet Network' is used for training in order to determine the coherent characteristics of a given sequence. X. Wang et al. (2016) also use a Siamese Network for learning of high-level descriptors and transformations. They split an action into three phases where, two are used for classification. The use of these types of methods is new to activity recognition, and requires further research.

Other methods

Beside the methods discussed earlier, there are also few methods available that models actions based on spatio-temporal features. For example, some researchers focused on textures for modeling actions. Among popular methods, Kellokumpu et al.

(2008a) proposed to use local binary pattern on three orthogonal planes (LBP-TOP) (G. Zhao & Pietikainen, 2007) (a formulation to capture LBP from three different planes, i.e. XY, XT, and YT) on an entire STV. Mattivi and Shao (2009) used Extended LBP-TOP – a formulation of LBP-TOP with nine intersection planes (each plane contributes three) based on Harris3D feature points. Kellokumpu et al. (2011) improved LBP-TOP by the use of human detection and hidden markov model (HMM). Baumann et al. (2014) used volume LBP (VLBP) for action modeling from STV, and in Baumann et al. (2016) they further extended it to motion binary pattern (MBP).

While most of the researchers concentrated on local STIP, some worked on the distribution of STIPs across the video. The local STIP methods are highly reliable on the discriminative capacities of feature descriptors, and denies the global distribution information of STIPs across the video which also has a potential of recognizing human activities. Bregonzio et al. (2009) proposed to use the global distribution of STIP (Dollár et al., 2005) for the description of human activities. X. Wu et al. (2011) used context and appearance distribution features for describing human activities. They represent each video as a set of 3D coordinates collected from interest points (IPs). Yuan et al. (2013) proposed 3D R (three dimensional Radon transform) to describe the distribution of STIPs across the activity video.

Besides these methods, efforts in other directions are also available such as action pose based histogram of oriented rectangles (Ikizler & Duygulu, 2009), hierarchical neighborhood features (Kovashka & Grauman, 2010), boosted eigen action (C. Liu & Yuen, 2010), body joint quadruples (Evangelidis et al., 2014), hierarchical ensemble tree (Ma, Sigal, & Sclaroff, 2015) and semantic feature fusion (J. Cai et al., 2015).

2.2 Low quality video-based approaches

There are very few methods available in the literature that addresses the problem of video quality, but instead focuses on low video resolution. Other quality problems such as compression, camera motion and blurring are relatively unexplored. In

Oh et al. (2011), the authors also state the same conclusions, and further emphasized in the development of methods for low quality videos. However, in this section, the methods that specifically address the problem of video quality or use the methodologies similar to ours are reviewed.

Among early methods, most works are over-reliant on motion information, as low resolution usually distorts important shape cues. Efros et al. (2003) first came up with idea of recognizing human activities from a distance. They proposed a motion based descriptor which is capable of encoding features from human figures that has a approximate height of 30 pixels. The descriptor is based on optical flow calculated from a smoothed human activity focused space-time volume. Lu and Little (2006) further extended their work by proposing PCA projected histogram of oriented gradients (PCA-HOG) feature for description of tracked video features. The usage of PCA on HOG improved the performance of feature tracking in video. However, these two methods are only sufficient for recognizing 'static' type of activities, where background of video is less complex such as KTH. In Efros et al. (2003), the computation of optical flow is activity centric (only calculates from the frames which has human), therefore for certain activities where consideration of related backgrounds is necessary, it may cause loss of motion information.

C.-C. Chen and Aggarwal (2009) used histogram of oriented gradients (HOG) and histogram of oriented optical flow (HOOF) features for describing successive time series of poses and movements respectively from overlapped STVs (an action video is divided into overlapped space-time volumes for extraction of efficient features). They have created a low resolution aerial dataset named UT-Tower, and tried to recognize human activities from it. To reduce the feature dimensionality and extraction of efficient features they have used supervised principal component analysis (SPCA).Their method can retain a very high result for low quality videos with significantly reduced feature dimensions (almost one fifth dimension of original features). However, calculating features from various overlapped STVs is computationally very expensive and may include many non-action related features. This method does not work well with

complex scenarios and reduction of feature dimensionality sometimes may cause loss of performance.

To inspire researchers on analyzing aerial imagery, Ryoo et al. (2010) organized a challenge named "Aerial View Activity Classification Challenge". They have organized the competition based on UT-Tower dataset. A team from the University of Boston participated with a method called 'action covariance manifolds' (Guo et al., 2010) and win the challenge by beating baseline method that consists of spatio-temporal histogram of oriented gradient features with linear support vector machine. The method action covariance manifolds represent an activity sequence as the shape of the silhouette tunnel (temporal sequence of local shape-deformations of centroid-centered object silhouettes). Each silhouette tunnel is represented by a thirteen-dimensional feature matrix. However, silhouette based methods are only suitable for simplistic videos because they mostly have less complex backgrounds and structure of object in video frames is easily differentiable.

C.-C. Chen and Aggarwal (2011) propose speech like mid level modeling of human activities in order to recognize them on low resolution videos (C.-C. Chen & Aggarwal, 2009). At first they detect spatio-temporal interest points (STIPs) from action video. A set of associated STIP detectors from every action category is learned by modified Adaboost algorithm in order to associate STIPs with actions. The boosted STIP detectors are then use to measure the likelihood of occurrence of local interest patterns from various body parts. The time series of likelihood occurrences are further fragmented into overlapped short time segments and then again transformed by 1D Fast Fourier Transformation (FFT) in order to incorporate an 'action spectrogram'. They achieved a relatively high accuracy across various low quality datasets includes KTH (90.9%), UT-Tower (98.2%) and VIRAT (Oh et al., 2011) (38.3%).

M. A. Ahad et al. (2010) extended the idea of motion history image (A. F. Bobick & Davis, 2001) and proposed a descriptor named directional motion history image (DMHI) for recognizing human activities form low-resolution videos. They used opti-

cal flow for computing DMHI images form low resolution video. They also presented a new low resolution dataset created by spatial downsampling, collected from various indoor scenarios. However, the assessment of their method for low quality videos is not much realistic since majority of the poor resolution videos are captured outside from a distance where air turbulence usually creates blurring effects.

Reddy et al. (2012) used histogram of 3D spatiotemporal gradients (3D-STHOG) to recognize human activities. They have also investigated the performance of 3D-STHOG under various quality conditions such as sampling rate, scale and rotation. They also reported that the features performance reported for small database does not always reflect the same if large-scale datasets is used. They also analysed minor shifting of visual data, which sometimes may lead towards a great performance loss, i.e. $\pm$ 2 shifting pixels in UT-Tower dataset cause 20% loss of performance. They also suggested to use a descriptor that has viewpoint handling capability in order to have a better performance.

Harjanto et al. (2015) investigates the impact of frame rates on four different activity recognition methods for robust recognition of human activities. They use a key-frame selector to select key frames from video. Their experimental results shows that selection of representative action frames from video helps to achieve a higher performance across a verity dataset. However, their key frame selector may not always applicable for videos where spatial resolution is low or poor. And for low quality videos, feature representation from a selective frames are not always enough to have a decent feature set.

2.3 Summary

In recent years, human activity recognition has received much attention due to its real-world application domain, especially from the pattern recognition and computer vision communities. The growth of this area is motivated by the growth of techniques in image based recognition (detectors and descriptors), feature learning and classification methods, and the methods with the intuition from psychology and

biology.

In this chapter, a structural review of existing HAR methods and critical analysis are presented. In order to summarize this review of existing works, the following remarks deserved mentioning:

1. Volume-based approaches are less popular in literature due to its major disadvantage of recognizing the actions involving multiple persons. There is an apparent difficulty in spatially segmenting the volume of actions belonging to persons in close proximity. Nevertheless, volume-based approaches offer some high-level representation of human actions (with minimal loss of information) across the temporal dimension, with elegant similarity matching techniques between volumetric patches.

2. While there is an increase of attention in the feature based spatio-temporal approaches due to their reliability under noise and illumination changes, the major limitation lies in its suitability for modeling more complex activities, not just simple periodic activities involving single person interactions only. Despite the ability to handle scale invariance well, feature-based approaches struggle to handle viewpoint invariance, especially in wide surveillance areas where persons or objects of interest can undergo a tremendous change of view angle as they move in the monitored scene.

3. On the other hand, trajectory-based spatio-temporal approaches are view-invariant, but this comes at a cost – low-level estimation of 3-D XYZ joint or body part locations of moving persons which is a difficult and expensive requirement for success. However, recent trajectory methods such as improved dense trajectories offer noise and camera motion free features that works well with many complex activity scenarios.

4. Unsupervised learned deep features perform well in big data scenarios. Convolutional Neural Networks (CNN) networks and their explicit types such as Multi-stream Networks are widely used methods for activity recognition, and performs bet-

ter than handcrafted methods for large-scale complex videos. However, training deep network is computationally very expensive.

5. There is no authoritative work in literature that directly address and propose techniques for activity recognition in quality video. A few related works are available, but they only address the problem of low resolution videos. Due to shape distortion, most methods used only motion features as an activity cue.

All the mention approaches are not directly applicable to the problem of human activity recognition in low quality videos. Therefore considerable further work is required to deal with this problem.

CHAPTER 3

METHODOLOGY AND DATASETS

In this chapter, research methodology, various datasets, their respective experimental protocols and evaluation matrices used in this thesis are described. Since the aim of this thesis is to recognize human activities from low quality videos, there is no specific dataset available in literature that directly addresses the issue of low quality video. For the purpose of this research, various low quality versions from two low quality publicly available datasets – KTH actions (Schüldt et al., 2004) and UCF-11 (J. Liu et al., 2009) are created. Another large-scale dataset, HMDB51 (Kuehne et al., 2011) provided quality meta-labels which were useful. Beside the evaluation strategies, the methodologies used for the creation of low quality videos from a higher quality dataset are also discussed.

3.1 Methodology and General Framework

We first describe the research methodology and general framework employed in this research. An overview of our research framework is shown in Figure 3.1. It comprises of several steps and a brief detail of each of them are as follows:

Input Video: We have used three well-known activity recognition datasets as discussed in Section 3.2. Since original datasets were created using only good quality videos so, we created low quality versions/subsets of them for use n our research. Only these new created versions/subsets were used for the purpose of training and testing throughout the research.

Feature Extraction: A set of features expressing the activity in video is extracted from both training and test videos. The feature extraction process typically comprises of two steps, feature *detection* and *representation*. In *detection* step, space-

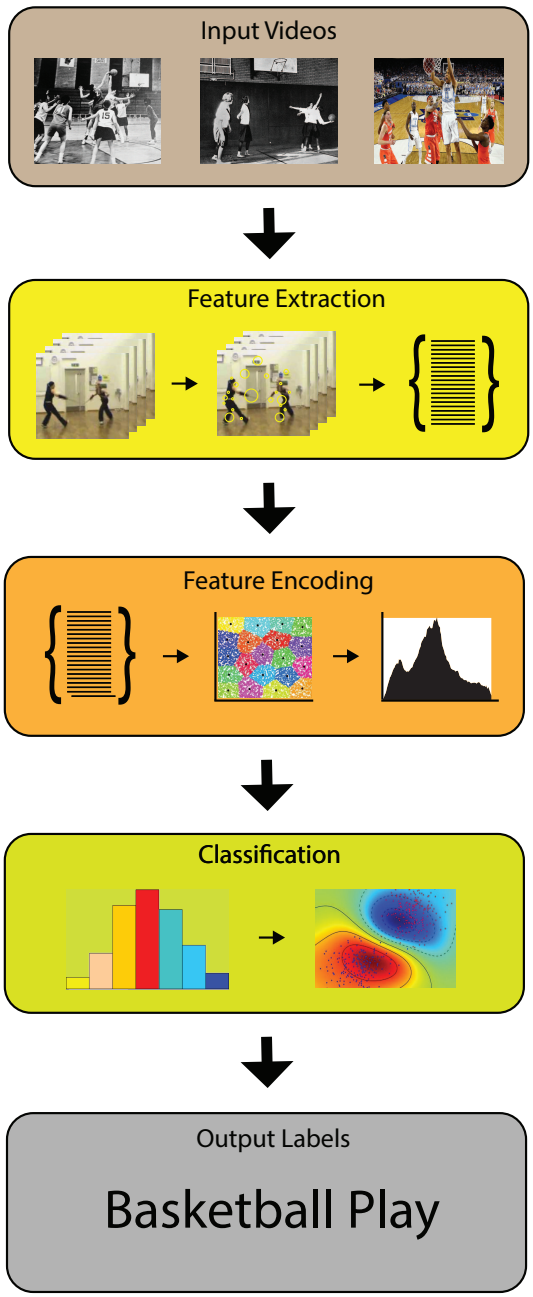


Figure 3.1: Overview of research methodology for human activity recognition in low quality videos

time interest points or trajectories are detected and in *representation* step, feature descriptors are used to describe the feature patterns surrounding the detected points or trajectories. This is only applicable if local features are extracted. In the case of global feature, extraction is performed on the whole frame and it typically skips the use of feature detectors.

Feature Encoding: In feature representation, all features extracted in the 'Feature Extraction' step are utilized to combine them into a holistic representation. There are many popular encoding techniques available, however, we only utilized *Vector Quantization* and *Fisher Vector*. Nevertheless, some global feature representation methods bypass this step since they produce the same number of features for each video sample.

Classification: A multi-class classifier is trained using features extracted and encoded from all training videos, it is used for classification of test videos.

Output Labels: Each test video is assigned with a predicted label upon classification, and the overall results are measured by performance metrics (see Section 3.3).

3.2 Datasets for evaluation

In this section, various datasets that we have used for assessing our proposed methods, namely the KTH action, the UCF-11 (formerly known as UCF-YouTube) and the HMDB51 are described. All these are considered benchmark datasets and are popularly used by the activity recognition community. The detailed descriptions of the aforementioned datasets and methodologies for creation of low quality videos are discussed in the following subsections.

3.2.1 KTH action dataset

KTH action (Schüldt et al., 2004) is considered as the most popular dataset in literature for human activity recognition. It has 6 types of human activity classes, namely *boxing, running, hand clapping, jogging, hand waving* and *walking*. There are a total of 25 different human actors acted in four different controlled environments to form this dataset. The environments or scenarios are indoors, outdoors, outdoors with dissimilar clothes and outdoors with the variations of scales. There are 599 video samples in total (one subject in *hand-clapping* action class has one clip less). Each clip is sampled at 25 fps (frames per second) and lasts between 10-15 seconds with

an image frame resolution of 160 $\times$120 pixels. Figure 3.2 shows some sample videos of KTH dataset representing each activity classes from four activity scenarios. We follow the As an original experimental setup as in Schüldt et al. (2004), where all video samples are split into two sets: a test set and a training set. The videos of subjects 2-3, 5-10, and 22 are declared as a test set and the remaining 16 subjects are declared as a training set. For evaluation, the average accuracy across all test classes is used as a measurement of performance. The videos in KTH offer relatively good spatial quality that gives clear visual information. In order to assess low quality videos spatial and temporal downsampling were applied on original videos. The effects of downsampling critically affect the quality of videos from two viewpoints: resolution and sampling rate. The detailed description of spatial and temporal downsampling is given in the section 3.2.1 (a) and 3.2.1 (b) respectively.

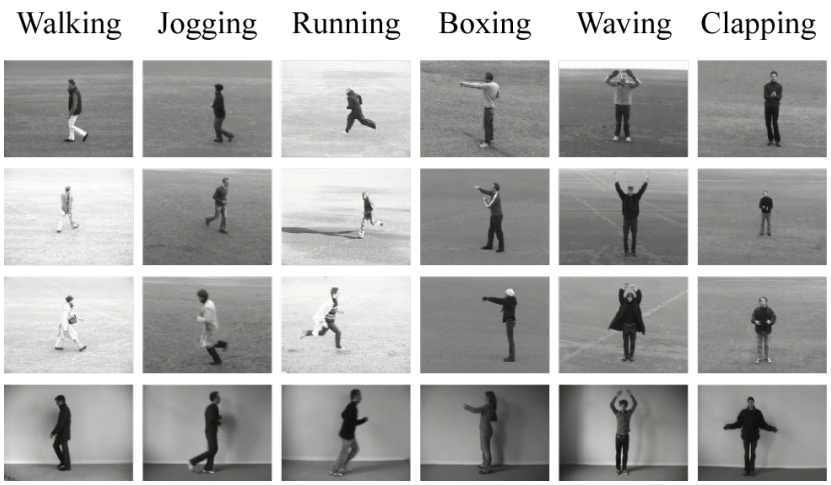


Figure 3.2: Sample video frames from KTH action dataset representing each action classes under four different scenarios

3.2.1 (a) Spatial Downsampling

Spatial downsampling (SD) produces an output video with a smaller resolution than the original video. In the process, no additional data compression is applied while the frame rates remained the same. For clarity, we define a spatial downsampling factor, α which indicates the factor in which the original spatial resolution is reduced. In this work, we fixed $\alpha = \{2, 3, 4\}$ for modes SD_α, denoting that the original videos are

to be downsampled to half, a third and a fourth of its original resolution respectively. Figure 3.3 shows a sample video frame that undergoes SD_1, SD_2 and SD_3. We opted not to go beyond $\alpha = 4$ as the extracted features are too few and sparse to provide any meaningful representation.

Figure 3.3: Spatially downsampled videos. (a) Original (SD_1), (b) SD_2, (c) SD_4, (d) SD_4; Figure reproduced from Rahman et al. (2015).

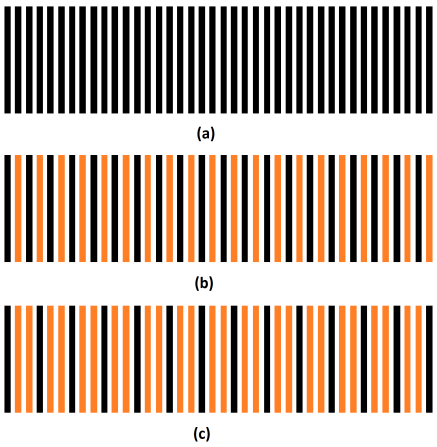


Figure 3.4: Temporally downsampled videos. (a) Original (TD_1), (b) TD_2, (c) TD_3; Figure reproduced from Rahman et al. (2015).

3.2.1 (b) Temporal Downsampling

Temporal downsampling (TD) produces an output video with smaller temporal sampling rate (or frame rate) than the original video. In the process, the video frame resolution remained the same. Likewise, a temporal downsampling factor, β is defined which indicates the factor in which the original frame rate is reduced. In this work, $\beta = \{2, 3, 4\}$ are used for modes TD_β, denoting that the original videos are to be

downsampled to half, a third and a fourth of its original frame rate respectively. Figure 3.4 shows how a video sequence frame undergoes temporal downsampling at equal intervals. Frames indicated by orange color are discarded while frames in black are kept. In the case of videos with slow frame rates or short video lengths, β may only take on a smaller range of values to ensure sufficient features can be extracted for representation.

3.2.2 UCF-11 dataset

The 'UCF-11', also known UCF-YouTube (J. Liu et al., 2009) is another popular dataset for activity recognition, consisting of videos from an uncontrolled and challenging environment. It contains 11 activity classes and every class is divided into twenty-five groups. Each group has minimum of four video clips, and every video clip in the same group have few common characteristics, such as similar actor, viewpoint, background etc. The videos in UCF-11 contained various problems such as camera motion, background clutter, viewpoint and scale variations. There are 1600 video samples in total and each clip is sampled at 30 *fps* at a frame resolution of 320×240 pixels. The sample videos from the UCF-11 dataset is shown in Figure 3.5. For evaluation, we use leave one group out cross validation (LOGOCV) scheme specified by J. Liu et al. (2009), and report average accuracy over all classes.

Since we are only interested in evaluating low quality videos, so we perform further compression on each video sample to make the dataset more challenging. Specifically, we re-encode all video samples by using x264 video encoder (Wiegand et al., 2003) based on a uniformly distributed Constant Rate Factor (CRF) values. These CRF values were randomly assigned to each video samples[1], but done in a manner which is uniform across a CRF range of 23 to 50. Higher CRF values lead to greater compression effects and smaller file sizes and vice versa. For clarity, we call this newly created version as **YouTube-LQ**. Some sample videos created with different CRF values are shown in figure 3.6. As we can see from the figure, the videos that

[1]The distribution of various CRF values across video samples can be obtained from http://saimunur.github.io/YouTube-LQ-CRFs.txt

Figure 3.5: Sample video frames from UCF-11 dataset. Videos from 8 action classes is presented. Figure reproduced from J. Liu et al. (2009).

have a higher CRF value are lack of detailed visual information (the shape information became distorted due to compression).

Figure 3.6: Sample video frame with different constant rate factor (CRF) values: CRF 29 (left), CRF 38 (center) and CRF 50 (right).

3.2.3 HMDB51 dataset

The HMDB51 dataset (Kuehne et al., 2011) has 6766 videos distributed over 51 activity classes. It is one of the largest dataset in human activity recognition. The videos in this dataset were collected from YouTube and digital movie clips. Video clips depict mainly natural actions from uncontrolled environments (i.e., “in the wild”), with a wide range of camera viewpoints, the presence of camera motion, involvement of

different number of humans in a particular activity. Each category contains at least 101 video clips. Beside the action labels, every video clip is also marked with meta-labels describing various properties of the clip including the quality of video. For each video clip, a three-level scheme was applied to grade the quality.

A criteria was set to gauge the ease of observers in identifying individual fingers in the video motion. Video samples that failed this criteria were labeled 'bad' or 'medium' depending on the visibility of limbs and parts of body during the activity. In addition, the 'bad' videos also contain significant motion blurring and compression artifacts. Figure 3.7 shows sample frames from 'bad' and 'medium' quality video clips of few selected action classes from the HMDB51 dataset.

Figure 3.7: Sample 'low' and 'medium' quality video clips from HMDB51.

For the purpose of this work, we are mainly interested in the evaluation of clips annotated with 'medium' and 'bad' quality labels. However, the 'high' quality clips are also useful as a control experiment for the sake of comparison. Hence, we partition the HMDB51 dataset into three subsets based on its quality label (distribution in parenthesis): HMDB51-BQ (20.8%) containing 'bad' quality clips, HMDB51-MQ (62.1%) containing 'medium' quality clips, and HMDB51-HQ (17.1%) comprising of the remaining 'high' quality clips. For consistency of experiments, we follow the settings used in the original paper (Kuehne et al., 2011) whereby three distinct training-

test splits (70/30 clip distribution per class) were used. The mean accuracy of all three splits is reported as the final measure of performance.

3.3 Performance Metrics

Standard benchmark metrics for human activity recognition are employed to gauge the performance of proposed methods. For all datasets, average accuracy across all tested classes is reported. Confusion matrices in specific cases are also necessary to observe class-specific performances. The following performance metrics are used for evaluations:

Accuracy: The accuracy or the 'hit' rate indicates the proportion of correctly predicted items over all possible test items from all test classes. The accuracy was calculated using the following:

$$\text{Accuracy} = \frac{\textit{Correctly predicted samples}}{\textit{All test samples}} \times 100\% \tag{3.1}$$

Confusion matrix (CM): A CM shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes. In CM, the columns represents the predicted class instances and the row represents actual class instances. Mean accuracy (across all classes) can be computed from all tested samples that belong to every tested class

3.4 Conclusion

In this chapter, various datasets used for assessing proposed methods are described. Due to the lack of low quality human activity dataset in literature, we have created various low quality versions or subsets from three existing publicly available datasets. The creation of these low quality videos is done in a systemic way to help evaluate proposed methods thoroughly. The performance evaluation metrics that are used in experiments also reported in this chapter.

CHAPTER 4

JOINT FEATURE UTILIZATION FOR ACTIVITY RECOGNITION IN LOW QUALITY VIDEO

Feature representation is a fundamental part of computer vision which has a great control over the performance (D. Zhao et al., 2013). With the growth of visual data such as image and video, feature representation plays an important role in various applications such as face recognition, emotion recognition, image classification and also in activity recognition, where an efficient representation of features is necessary due to challenging nature of the data.

Generally, an action in a video can be represented in a form of vectors or matrices (by use of feature descriptors). However, the representation is solely based on feature descriptor and the efficiency of features depends on how well the method is designed to encode the action. Current frameworks use shape and motion features to represent action which are not appropriate for low quality videos due to the complexity of visual data. Hence, efficiency of current frameworks can be improved by extracting statistical regularity feature between video frames.

In this chapter, a joint feature utilization framework or method to better characterize activities in low quality videos is proposed. The framework extracts multiple spatio-temporal video features and combine them in an efficient way that increase the robustness of activity classification in lower quality videos. The conception of combing features is inspired by recent activity recognition methods such as Laptev (2005), Y. Wang and Mori (2009) and H. Wang and Schmid (2013). These methods show that use of multiple features can increase the performance of classification. In addition to the performance, the use of multiple features helps to improve the discriminative capacity of the overall feature representation.

For each action video, a set of features, i.e. shape, motion and texture, are extracted. The shape and motion features are represented as a histogram of features by using bag-of-visual-word (BoVW) framework. The BoVW represent each video as a histogram of features by following three distinct steps, namely, codebook generation, encoding, and pooling-normalization (a detailed description of each BoVW step is given in Appendix 1). The textural feature histograms are concatenated to increase the robustness of features. Finally, test videos are classified using a non-linear support vector machine (SVM). The performance of joint feature utilization was evaluated with low quality versions and subsets of various publicly available datasets.

Section 4.1 describes the related methods and motivations behind the idea of joint feature utilization framework. Various shape, motion and textural feature based methods available in the literature are thoroughly discussed and their suitability for low quality videos are critically analyzed. Section 4.2 describes various spatio-temporal feature representation methods that are used in experiments. Section 4.3 describes the joint feature utilization framework and its various components. Section 4.4 shows the experimental results with further analysis. Finally, section 4.5 gives a summary of this chapter.

4.1 Related Work and Motivations

In recent years, many activity recognition methods have been proposed in literature. Generally, feature representation can be classified into two types, namely local and global representation. Local feature based methods represent human activities in terms of key poses or parts from a video that has significant action changes while global methods considers whole video frames to represent human activities. Shape, motion and texture are the most widely used features for human activity recognition. Among them, shape and motion have shown excellent performance on various challenging datasets that comprises of many complex activity recognition problems such as illumination variations, clutters and camera motions. However, each has their individual limitations that becomes more apparent when video quality becomes poor. For example, shape and motion features that are based on gradients such as HOG and HOF

(Laptev et al., 2008) do not provide significant orientation changes (even sometimes no changes) if video quality becomes poor. Similarly, texture features such as LBP (Ojala et al., 2002) and BSIF (Kannala & Rahtu, 2012) also do not offer good performance if video quality becomes poor.

In recent activity recognition methods, many researchers have shown that use of multiple features brings about better performance than using single features (Laptev et al., 2008; H. Wang et al., 2009, 2011; D. Zhao et al., 2013; H. Wang & Schmid, 2013; Peng et al., 2014). It also provides robustness towards many complex activity recognition problems. Many methods or frameworks have been proposed in literature to utilize multiple features for activity recognition (D. Zhao et al., 2013). Based on the nature of representation, they can be divided into two types, namely (1) shape-motion features and (2) textural features. The details of both feature types and their variants are given below:

4.1.1 Shape and motion features and their variants

Local shape and motion feature based methods (Laptev, 2005; Dollár et al., 2005; H. Wang et al., 2009, 2011; H. Wang & Schmid, 2013) have become very popular due to their excellent results in activity recognition. Many of these methods appear to suggest that use of multiple feature in a complementary fashion achieves higher performance than using them alone. Recent works on feature fusions (D. Zhao et al., 2013; Peng et al., 2014; L. Wang et al., 2014) also suggests the same.

Many of the early activity recognition methods were based on spatio-temporal interest points. One of the most popular method by Laptev et al. (2008) combined appearance and motion features to achieve higher performance in recognizing activities from movies. They have used Harris3D (Laptev, 2005) detector to detect space-time interest points (STIPs), and HOG and HOF feature descriptors across those points to represent shape and motion features. They achieved 91.8% accuracy on the KTH action dataset (Schüldt et al., 2004) using the traditional BoVW feature encoding. However, gradient based descriptors such as HOG and HOF are sensitive to illumination

changes, and the quantization of features does not consider any relationships between the spatial and temporal domain. Klaser et al. (2008) generalizes the concept of 2D gradient descriptors such as HOG to HOG3D. They used regular polyhedrons to encode gradient orientations, which showed improvement of results in comparison with the regular HOG (Laptev et al., 2008) (for KTH, the performance of HOG3D was 91.4% and HOG was 81.6%).

The Harris3D detector offers only a number of interest points that may not be always useful if video is captured in a complex environment. In order to achieve more dense space-time interest points from realistic videos captured from complex environments, J. Liu et al. (2009) used Cuboid detector (Dollár et al., 2005) to detect interest points (IPs). Another popular detector is the Hessian detector (Willems et al., 2008), which produces dense scale invariant interest points. H. Wang et al. (2009) evaluated various detectors including Hessian and cuboid detectors on various feature descriptors such as ESURF, HOG3D, HOG and HOF. They also evaluated the combined HOG and HOF features, and demonstrated that the combination of HOG and HOF outperforms single descriptors (descriptors that are not concatenated with any other) and significantly improves the performance by a good margin across various datasets including KTH. However, in an environment where videos are captured in a controlled scenario, dense detectors did not perform well. Rather, it performed better with videos captured from complex environments.

A recent work by D. Zhao et al. (2013) on combining appearance and structural features also obtained improved results. While most works only focused on the appearance features, they also encode structural features using an optimized 3D shape context descriptor across the interest points (Dollár et al., 2005) and combined them with appearance features. However, their method does not work well with complex videos and offers a relatively poorer performance than current state-of-the-art methods such as STIP (Laptev et al., 2008).

Besides interest point based features, densely sampled (H. Wang et al., 2009)

features are also popular in literature. The space-time feature blocks at regular position and scales were extracted from the videos, and represented using various feature descriptors. HOG, HOF and HOG3D feature descriptors were evaluated across publicly available datasets, and showed that the combined use of shape and motion features, i.e. HOG/HOF, outperforms their individual use. However, dense sampling is computationally expensive and does not work well with videos captured in a controlled environment.

The use of trajectories has also recently become popular for their excellent results in activity recognition. H. Wang et al. (2011) proposed an optical flow based tracking of densely sampled points to obtain the trajectories. The visual patterns across the trajectories were described using various descriptors such as HOG, HOF and MBH. They also demonstrated that the use of multiple features greatly improves the recognition performance. Dense trajectories offer a relatively good performance when videos with complex background are considered. However, the method still suffers from camera motion and relative noise. Their subsequent work called improved dense trajectories (H. Wang & Schmid, 2013) improves these problems by extracting more efficient trajectories from video. They used warped flow estimation to remove irrelevant background motions from video and describe them with the same feature descriptors as before. The experimental results of this method also show that the use of multiple features greatly helps to achieve better performance across various datasets.

4.1.2 Textural features and their variants

The use of textural features (Kellokumpu et al., 2008b, 2008a; Mattivi & Shao, 2009; Kellokumpu et al., 2011; Ahsan et al., 2014; Baumann et al., 2016) is less common in literature of activity recognition. Although, their reported results were promising, but they are still relatively lower than the performance of shape and motion features.

Kellokumpu et al. (2008a) extended the idea of using local textural features on spatio-temporal domain (Kellokumpu et al., 2008b). They used local binary pat-

tern (LBP) features on three orthogonal planes (LBP-TOP) (G. Zhao & Pietikainen, 2007) to represent action video as a form of dynamic textures. The combination of textural features from three orthogonal planes greatly helps to improve the recognition performance. Their proposed method was capable of capturing shapes and space-time transition information with less computational complexity. However, due to its holistic nature, their method is easily affected by unnecessary background information and occlusions.

Unlike global texture representation, extraction of textures from local IPs is also found in literature. Mattivi and Shao (2009) proposed a method that uses part based representations such as interest points to overcome background and occlusion related problems. They used the Dollár detector (Dollár et al., 2005) to extract cuboids from video, and describe them using Extended LBP-TOP descriptor (an extension of LBP-TOP to nine slices, three for each plane). They have also demonstrated that the combined use of LBP from many planes, i.e. nine planes, offers better performance than using features from less planes, i.e. three planes. Their proposed formulation of LBP captures more motion information than the implementation of G. Zhao and Pietikainen (2007), but it also increases the computational complexity.

Beside the straightforward use of LBP on video frames, efforts in other directions, such as the use of LBP to describe feature templates are also found in literature. Kellokumpu et al. (2011) used LBP features to describe motion history images (MHI) and motion energy images (MEI) (A. Bobick & Davis, 1996) to receptively encode shape and motion information. They have managed to achieve a good performance on the KTH (Schüldt et al., 2004) and Weizmann (Blank et al., 2005) datasets by utilizing these features in a combined manner. Ahsan et al. (2014) also used LBP features to describe mixed block based directional MHI (DMHI) templates (M. A. R. Ahad et al., 2008). Like Kellokumpu et al. (2011), they have also experienced that the joint use of MHI images from different optical flow directions improves the recognition performance. However, the problem is, LBP based methods are sensitive to noise and illumination changes (Yeffet & Wolf, 2009), and lack of explicit motion encoding.

There are also few works available related to the evaluation of textural features. A recent work by Kataoka et al. (2015b) used dense trajectory estimated shape, motion and texture features, and evaluated their individual and combined performances for activity recognition. In evaluation, they found that the joint use of these features helps to improve the recognition performance across various realistic video datasets. They also observed that the textural features do not offer better performance than shape and motion features, instead it helps these features to improve their performance.

Motivated by the analysis above, this chapter proposes a joint feature utilization framework that extracts various spatio-temporal shape-motion and textural features, and jointly utilizes them to recognize human activities from low quality videos. While many of the existing frameworks only use conventional shape and motion features, this framework also utilizes textural features to vastly improve the recognition of human activities under low quality conditions.

4.2 Spatio-Temporal Features

This section describes various spatio-temporal features including their detection and description process used in the joint feature utilization framework.

4.2.1 Shape and Motion Features

Shape and motion are an expressive abstraction of visual patterns in action video. Shape and motion are critical cues for recognition, as they are sufficiently invariant to represent commonalities of different instances of a particular action category, while preserving enough detail of actions in order to differentiate them from each other. As our main goal is to obtain robust shape and motion features, we detect them based on two recently proposed methods, namely space-time interest points (Laptev, 2005) and improved trajectories (H. Wang & Schmid, 2013), and describe them with popular gradient based feature descriptors such as the histogram of oriented gradients (HOG), histogram of optical flow (HOF) (Schüldt et al., 2004) and motion boundary histogram (MBH) (H. Wang & Schmid, 2013). We further process these feature descriptors using bag-of-visual-words (BoVW) method (X. Wang et al., 2013) and represent each

action video as a histogram of feature occurrences. A brief description of these feature detection and description methods are given as follows:

a) Space-time interest points: Given an action video, local space-time interest points (STIP) are detected around the location of large variations of image values, i.e. motions. In order to detect interest points, the Harris3D detector (Laptev, 2005) is used. Harris3D detector is an extension of the popular Harris detector widely used in the image domain (Harris & Stephens, 1988)) for corner detection. It can detect a decent amount of corner points in space-time domain and is perhaps one of the most widely used feature detector for action recognition. In brief, a second moment space-time matrix at each video location μ described as:

$$\mu(.;\sigma;\tau) = g(.;s\sigma;s\tau) * (\nabla L(.;\sigma;\tau)L(.;\sigma;\tau))^T \tag{4.1}$$

where, μ and τ are spatial and temporal scales, g is a separable Gaussian smoothing function and ∇L is the spatio-temporal gradients. The location of STIPs are determined by the local maxima of H calculated as:

$$H = det(\mu - \mathrm{k}trace^3(\mu)), H > 0 \tag{4.2}$$

Instead of using points from a particular scale, points extracted at multiple scales based on regular sampling of scale parameters α and τ are used. To characterize the shape and motion information accumulated in space-time neighborhoods of the detected STIPs, the Histogram of Gradient (HOG) and Histogram of Optical Flow (HOF) feature descriptors as proposed in (Schüldt et al., 2004) are used. The combination of HOG/HOF descriptors produces descriptors of size $\Delta_x(\sigma) = \Delta_y(\sigma) = 18\sigma$, $\Delta_t(\tau) = 8\tau$ (σ and τ are the spatial and temporal scales). Each volume is subdivided into $n_x \times n_y \times n_t$ grid of cells; for each cell, 4-bin histograms of gradient orientations (HOG) and 5-bin histograms of optical flow (HOF) are computed. The original implementation from (Laptev, 2005) and standard parameter settings from (H. Wang et al., 2009), i.e. k=0.0005, σ^2={4,8,16,32,64,128}, τ^2={2,4}, $\{n_x, n_y\}$=3 and $n_t = 2$ are

used in our experiments.

b) Space-time trajectories: Given an action video, space-time trajectory captures motion information by sampling interest points at a uniform interval and tracking them over a fixed number of frames. To detect trajectories, improved dense trajectories (iDT) (H. Wang & Schmid, 2013), which is an extension of dense trajectories (H. Wang et al., 2011) is used. At first, a set of points is densely sampled on a grid of 8 different spatial scales with a step size of 5 pixels. Points from homogeneous areas are removed by thresholding small eigenvalues of their respective autocorrelation matrices. Tracking of these sampled points is then done by applying median filtering to a dense optical flow field. To be precise, points on patch $P_t = (x_t, y_t)$ at frame t is tracked to another patch P_{t+1} in the next frame defined as:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t + y_t) + (M \times \omega_t)|_{(\bar{x}_t, \bar{y}_t)} \tag{4.3}$$

where, M is kernel for median filtering, $\omega_t = (u_t, v_t)$ is the dense optical flow field from $t+1$ frame, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of (x_t, y_t). In order to avoid drifting problem in tracking, trajectory length L is set to a constant value. At last, static trajectories with lack of motion information and trajectories with large displacement due to incorrect optical flow calculation were removed.

Improved trajectories are capable of boosting recognition performance by considering camera motions in action videos. It characterizes background motions between two consecutive frames by estimating homography matrix. To calculate the homography matrix, it first finds the similarities between two consecutive frames. Then, it uses SURF (H. Wang et al., 2009) and optical flow based feature matching since, they are complementary to each other. After finding feature similarities, it applies RANSAC (Fischler & Bolles, 1981) algorithm to calculate homography matrix. Based on that, they remove camera motion from video frame and re-compute the optical flow; that known as *warped flow*. The calculation of warped flow helps descriptors such as HOF and MBH to have a better motion estimation, free from camera motions.

In this work, iDT is used with a modification. It has been observed that tracking points on multiple spatial scale is computationally very expensive and time consuming. So, only tracked points on the original spatial scale is used to extract features. This is fast for implementation and still offers a decent recognition rate (2-3% less than the multi-scale extensions (H. Wang et al., 2013)). In brief, given an action video V, N number of trajectories is obtained:

$$\mathscr{T}(V) = \{T_1, T_2, T_3, ..., T_N\} \tag{4.4}$$

And T_n is the n^{th} trajectory at original spatial scale:

$$T_n = \{(x_1^n, y_1^n, t_1^n), (x_2^n, y_2^n, t_2^n), (x_3^n, y_3^n, t_3^n), ..., (x_p^n, y_p^n, t_p^n)\} \tag{4.5}$$

where P is trajectory length and (x_p^n, y_p^n, t_p^n) is the p^{th} position of the pixel at trajectory T_n. Only these trajectories will be considered for feature extraction from videos.

In this work, only motion boundary histogram (MBH) descriptor is used to describe features from detected trajectories. Like the HOF descriptor, MBH uses optical flow information, but splits it into horizontal and vertical derivatives. Also, it is robust to the camera and background motions and shows superiority of results over other gradient based descriptors such as HOG and HOF. The MBH produces descriptor of size $N \times N \times L$, where N is the size of the space-time volume in pixels and L is the length of the trajectory. Each volume is then subdivided into $n_x \times n_y \times n_t$ grid of cells; for each cell, 8-bin motion boundary histogram in horizontal direction (MBHx) and 8-bin motion boundary histogram in vertical direction (MBHy) are computed. The original implementation from (H. Wang & Schmid, 2013) and standard parameter settings, i.e. L =15, W=5, N=32, $\{n_x, n_y\}$=2, n_t=3 are used in this work.

4.2.2 Textural Features

Three types of textural features are used in our experiments, namely, local binary pattern (LBP) (Ojala et al., 2002), local phase quantization (LPQ) (Päivärinta et al., 2011) and binarized statistical image features (BSIF) (Kannala & Rahtu, 2012). At

first, the estimation process of individual textural features are briefly described, then their representation method is explained.

a) LBP features: Local binary patterns (LBP) (Ojala et al., 2002) use binary patterns calculated over a region for describing textural properties of an image. The LBP operator describes each image pixel by relative gray levels of its neighboring pixels. If the gray level of the neighboring pixel is higher or equal, the value is set to one, otherwise zero. The descriptor describes the result over the neighborhood as a binary number (binary pattern):

$$LBP_{P,R}(x,y) = \sum_{i=0}^{N-1} s(n_i - n_c)2^i,\ s_x = \begin{cases} 1 & x \leq 0 \\ 0 & otherwise \end{cases} \tag{4.6}$$

where n_c corresponds to the gray level of the center pixel of a local neighborhood and N equally spaced pixels n_i on a circle of radius R. The $LBP_{P,R}$ operator produces 2^P output values, corresponding to the 2^P binary patterns that can be formed by the P pixels in the neighborhood set. The final feature histogram of a particular image is produced by computing the occurrence of its different LBP output values.

b) LPQ features: Local phase quantization operator (Päivärinta et al., 2011) uses the local phase information to produce blur-invariant image features extracted by computing short term Fourier transform (STFT) in rectangular neighborhoods N_x, which are defined as:

$$F(u,x) = \sum_{y \in N_x} f(x-y)e^{-j2\rho u^T y} = \mathbf{W}_u^T \mathbf{f}_x \tag{4.7}$$

where W_u is the basis vector and f_x is image samples across N_x. Four complex coefficients corresponding to 2D frequencies are considered in forming LPQ features: $u_1 = [a,0]^T$, $u_2 = [0,a]^T$, $u_3 = [a,a]^T$ and $u_4 = [a,-a]^T$, where a is a scalar. To express the phase information, a binary coefficient is then formed from the sign of imaginary and real part of these Fourier coefficients. An image is then produced by representing 8 binary values (obtained from binary coefficient) as the integer value between 0 and

255. Finally, an LPQ feature histogram is then constructed from the produced image.

c) BSIF features: Binarized statistical image features (BSIF) (Kannala & Rahtu, 2012) is a recently proposed method that efficiently encodes texture information, in a similar vein to earlier methods that produce binary codes (Ojala et al., 2002; Ojansivu & Heikkilä, 2008). Given an image X of size $p \times p$, BSIF applies a linear filter F_i learnt from natural images through independent component analysis (ICA), on the pixel values of X and obtained the filter response,

$$r_i = \sum_{u,v} F_i(u,v)X(u,v) = \mathbf{f}_i^T \mathbf{x} \tag{4.8}$$

where $\mathbf{f}$ and $\mathbf{x}$ are the vectorized form of F_i and W respectively. The binarized feature b_i is then obtained by thresholding r_i at the level zero, i.e. $b_i = 1$ if $r_i > 0$ and $b_i = 0$ otherwise. The decomposition of the filter mask F_i allows the independent components or basis vectors to be learnt by ICA. Succinctly, we can learn n number of $l \times l$ linear filters W_i, stacked into a matrix $\mathbf{W}$ such that all responses can be efficiently computed by $\mathbf{s} = \mathbf{W}\mathbf{x}$. Consequently, an n-bit binary code is produced for each pixel, which then builds the feature histogram for the image.

Spatio-temporal extension of textural features: Motivated by the success of recent works related to the recognition of dynamic sequences (Mattivi & Shao, 2009; G. Zhao & Pietikainen, 2007), we consider three orthogonal planes (TOP) approach to extend the textural operators to extract the dynamic textures. Given a video (XYT), the TOP approach extracts the texture descriptors along the XY, XT and YT orthogonal planes, where the XY plane encodes structural information while XT and YT planes encode space-time transitional information. The histograms of all three planes are concatenated to form the final feature histogram. Given a volumetric space of $X \times Y \times T$, the textural histogram can be defined as:

$$h_j^{plane} = \sum_{p \in plane} \mathscr{I}\{b_i(p) = j\} \tag{4.9}$$

where $j \in \{1,...,2^n\}$, p is a pixel at location (x,y,t) at a particular plane, and $\mathscr{I}\{.\}$

a function indicating 1 if true and 0 otherwise. The histogram bins of each plane are then normalized to get a coherent description, $\tilde{h}^{plane} = \{\tilde{h}_1^{plane}, ..., \tilde{h}_{2^n}^{plane}\}$. Finally, we concatenate the histograms of all three planes,

$$H = \{\tilde{h}^{XY}, \tilde{h}^{XT}, \tilde{h}^{YT}\} \tag{4.10}$$

The notation of LBP operator that calculates LBP from three orthogonal planes is defined as:

$$LBP - TOP_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T} \tag{4.11}$$

where $P_{XY}, P_{XT}, P_{YT}, R_X, R_Y, R_T$ denotes a neighborhood of P points equally sampled on a circle of radius R on XY, XT and YT planes respectively.

The notation of LPQ operator that calculates LPQ values from three orthogonal planes can be defined as:

$$LPQ - TOP_{W_x,W_y,W_t} \tag{4.12}$$

where the notation W_x, W_y, W_t denotes the rectangular neighborhood at each pixel position on XY, XT and YT planes respectively.

Finally, the notation of BSIF operator that calculates LPQ values from three orthogonal planes is defined as:

$$BSIF - TOP_{l,n} \tag{4.13}$$

where l denotes the size of rectangular filters and n is the representation bit size at each pixel position on XY, XT and YT planes respectively.

In this work, the parameter settings of LBP-TOP$_{8,8,8,2,2,2}$ with non-uniform patterns as suggested in Mattivi and Shao (2009), LPQ-TOP$_{5,5,5}$ as specified in Päivärinta et al. (2011) were used. Meanwhile, BSIF-TOP$_{9,12}$ was selected empirically for various datasets used.

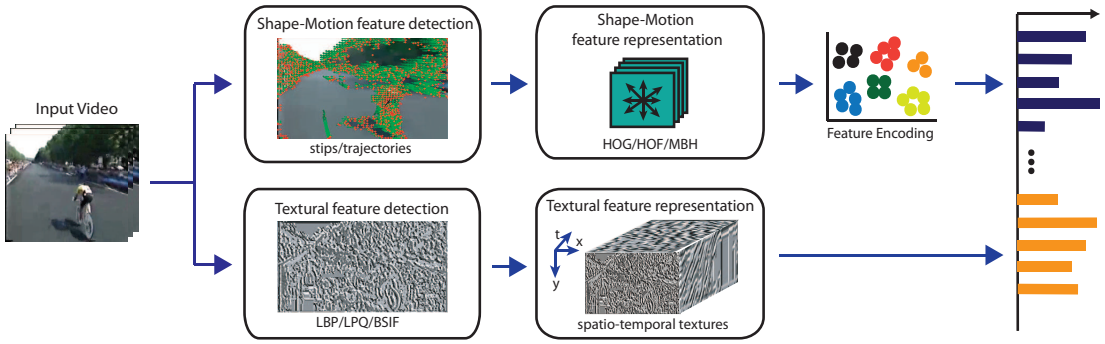


Figure 4.1: Proposed joint feature utilization based action recognition framework

4.3 Joint Feature Utilization Framework

In this section, a proposed framework for joint feature utilization for activity recognition in low quality video is described. The framework is shown in Figure 4.1. The main aim of proposed framework is to utilize textural information with conventional shape and motion features for better recognition of human activities in low quality videos. Every input videos goes through a series of steps. At first, space-time shape-motion features (interest point or dense trajectory based) are first detected and then represented by their respective descriptors. On the other hand, textural features (LBP, LPQ or BSIF features) are represented using three orthogonal planes (TOP) approach. The shape and motion features are encoded by bag-of-visual-words – vector quantization (VQ) (X. Wang et al., 2013) in order to obtain histogram level representation (the details of VQ is given in Section A.1 and A.2 of Appendix 1). Finally, both shape-motion and textural features are concatenated for classification by a non-linear multi-class support vector machine (SVM) with a chi-squared homogeneous kernel (Vedaldi & Zisserman, 2012), adopting a one-versus-all strategy. This homogeneous kernel is computationally efficient and provide the flexibility of deciding which features are to be 'kernelized' before a SVM classification.

While many frameworks (Laptev, 2005; H. Wang et al., 2009; H. Wang & Schmid, 2013; L. Wang et al., 2014) only extracts shape and motion features, proposed joint feature utilization framework also extracts textural features and combine them with shape and motion features in order to minimize their individual shortcomings.

In low quality videos where image quality and motion dynamics (transition of consecutive frames) is not good, extraction of robust shape and motion features (though selection of spatio-temporal interest points or trajectories) for action encoding is quite difficult. However, in such situation, statistical regularity of video frames, i.e. global textures is quite apparent and can be utilized to alleviate the limitations of shape and motion features.

4.4 Experimental Results and Analysis

In this section, experimental results for different downsampled videos of KTH, compressed videos of YouTube, and low and medium quality subsets of HMDB51 dataset are presented. In all experiments, the HOG, HOF descriptors are concatenate at the histogram level, denoted by 'STIP' and the concatenation of MBHx and MBHy descriptors is denoted by 'iDT'. Histogram level concatenation is found to be more effective than descriptor level concatenation (usually referred as HOG/HOF (H. Wang et al., 2009)) in our experiments. In the meantime, feature histograms from various textural features such as LBP-TOP, LPQ-TOP and BSIF-TOP, are extracted from entire video volume, and then concatenated with the STIP or iDT features. Throughout the experiments, STIP, and iDT features are considered as a baseline, with an intention to demonstrate that methods that incorporated textural features show significant improvement, as compared to baseline.

Experiments on downsampled videos: This section presents various experimental results on six downsampled versions of KTH action dataset. From the results reported in Table 4.1 (STIP based methods) and Table 4.2 (iDT based methods), methods that exploit additional textural features clearly demonstrate significant improvement, as compared to baseline methods. This is almost consistent across all six downsampled videos of KTH dataset except iDT based methods for TD_2 videos. The methods that use iDT features outperform STIP based methods across all downsampled versions. Among the textural features evaluated, BSIF-TOP appears to be the most promising choice, as it outperforms other textural features. Moreover, the contribution of textural features becomes more significant as the video quality deteriorates

(more noticeable in case of SD_4 and TD_4). This shows that textural features are more robust towards poorer video quality.

From the results for both STIP and IDT features, it is noticeable that the performance decrement is most obvious when dealing with spatial resolution. The feature detection from image frames is based on the variation of intensities in local image structures. The deterioration of spatial resolution critically affects the image intensities which results in failure to detect of important image features. In comparison to STIP, the performance of iDT features dropped by a great extent, especially for SD_3 and SD_4 videos.

Table 4.1: Recognition accuracy (%) of various STIP based feature combinations on downsampled versions of the KTH dataset.

Method	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
STIP **(Baseline)**	86.85	80.37	75.56	88.24	82.31	78.98
STIP+LBP-TOP	85.19	82.04	77.59	88.43	82.41	81.20
STIP+LPQ-TOP	87.41	80.19	76.30	87.41	81.85	79.81
STIP+BSIF-TOP	**88.80**	**85.28**	**81.67**	**88.70**	**86.11**	**84.54**

Table 4.2: Recognition accuracy (%) of various trajectory based feature combinations on downsampled versions of the KTH dataset.

Method	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
iDT **(Baseline)**	92.59	78.80	61.85	95.19	91.57	89.54
iDT+LBP-TOP	92.96	81.94	73.61	95.09	92.13	89.54
iDT+LPQ-TOP	92.96	78.61	79.91	95.09	91.67	88.89
iDT+BSIF-TOP	**93.89**	**88.33**	**82.41**	95.09	**92.22**	**90.00**

Figure 4.2 gives a closer look on detected features when videos are downsampled spatially and temporally. The videos on the first row are using Harris3D detector and videos on the second row are using warped flow estimation based feature tracking. The videos in column 1, 2, 3 respectively represent baseline, half resolution of baseline and one third resolution of baseline, and column 4 and 5 respectively represents

the half and one third frame rate of baseline. However, the spatio-temporal textures that rely on image regularity statistics, performs better in this kind of situations, but since it offers very poor performance on its own, so it is combined with STIP and iDT features. The textural features alone does not offer good performance but it does help other features to increase their performance (Kataoka et al., 2015b). For instance, in the case of STIP features, BSIF-TOP help to minimize the drop in accuracy by ≈6% for both SD_4 & TD_4 videos, and for iDT features, it minimizes ≈21% for SD_4 & ≈ 0.5% for TD_4 videos. The analysis of individual performance of various textural features is further discussed in the section related to performance analysis of individual textural features.

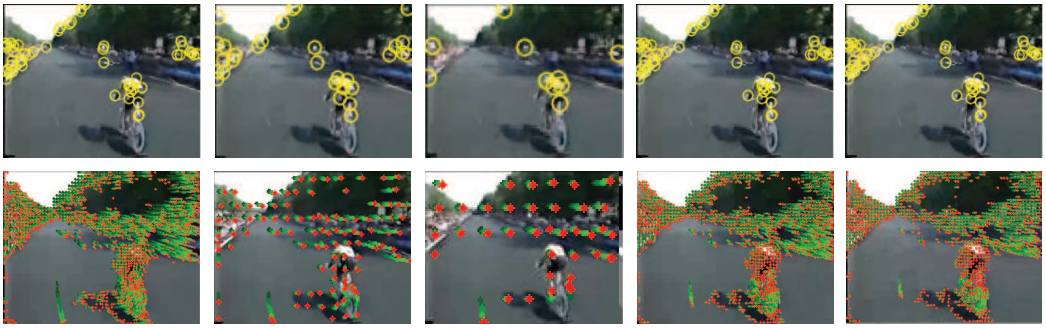


Figure 4.2: Response of detectors when videos are downsampled spatially and temporally (all videos are resized to same resolution for visualization). The sample video was taken from UCF-11 dataset (J. Liu et al., 2009).

Among various textural featured evaluated on both STIP and iDT features, BSIF-TOP appears to be the most promising choice, as it outperforms the others. With the degradation of spatial resolution and temporal sampling rate, BSIF-TOP comparatively performs better than LBP-TOP and LPQ-TOP features. Figure 4.4 and 4.3 gives a closer look at the performance improvement of BSIF-TOP features in comparison with LBP-TOP and LPQ-TOP. For instance, on STIP features, compared to LPQ-TOP, the improvement of BSIF-TOP is ≈5% for SD_4 and ≈4.8% TD_4 videos. On iDT features, the improvement of BSIF-TOP over LPQ-TOP is ≈10% for SD_3 and ≈0.5% for TD_3 videos. The performance improvement by LBP-TOP is a bit higher than the LPQ-TOP on both STIP and iDT features, except for iDT features of SD_4 videos.

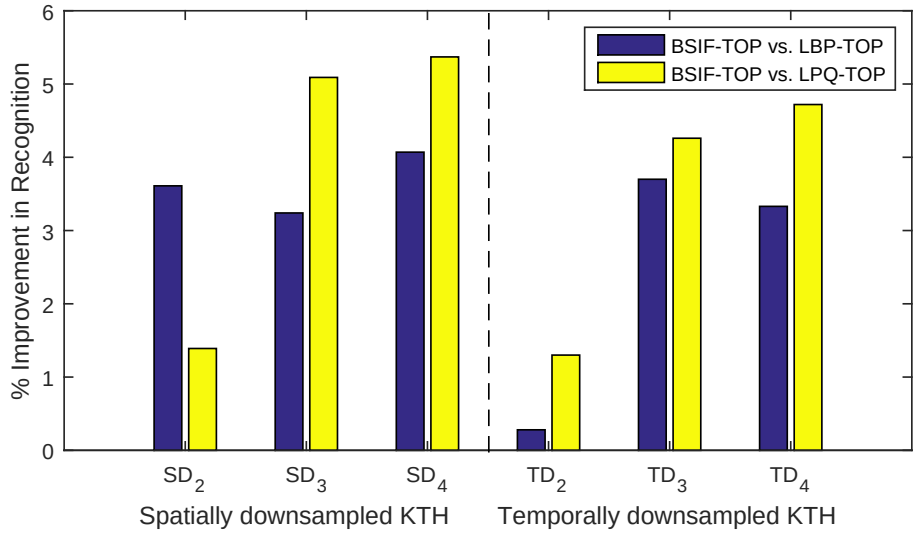


Figure 4.3: Percentage improvement of BSIF-TOP over LBP-TOP and LPQ-TOP, when combined with STIP.

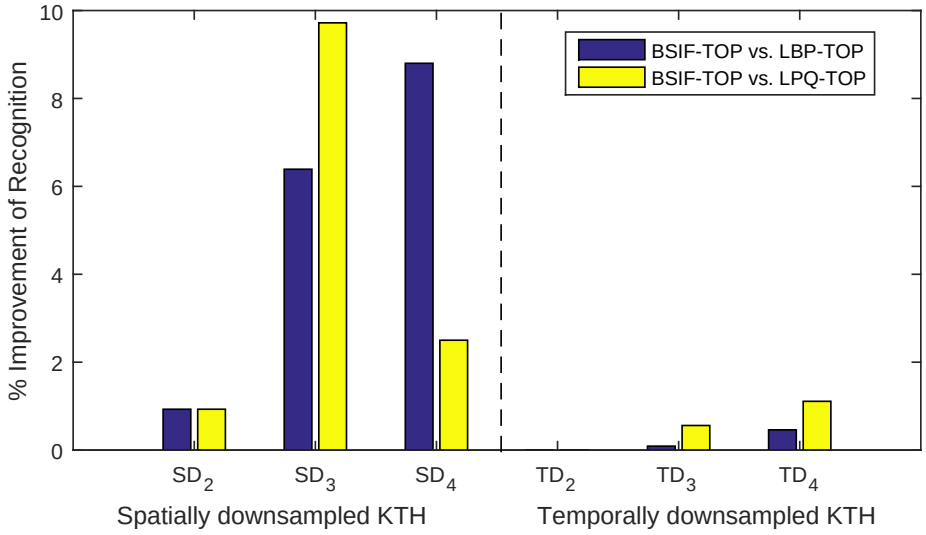
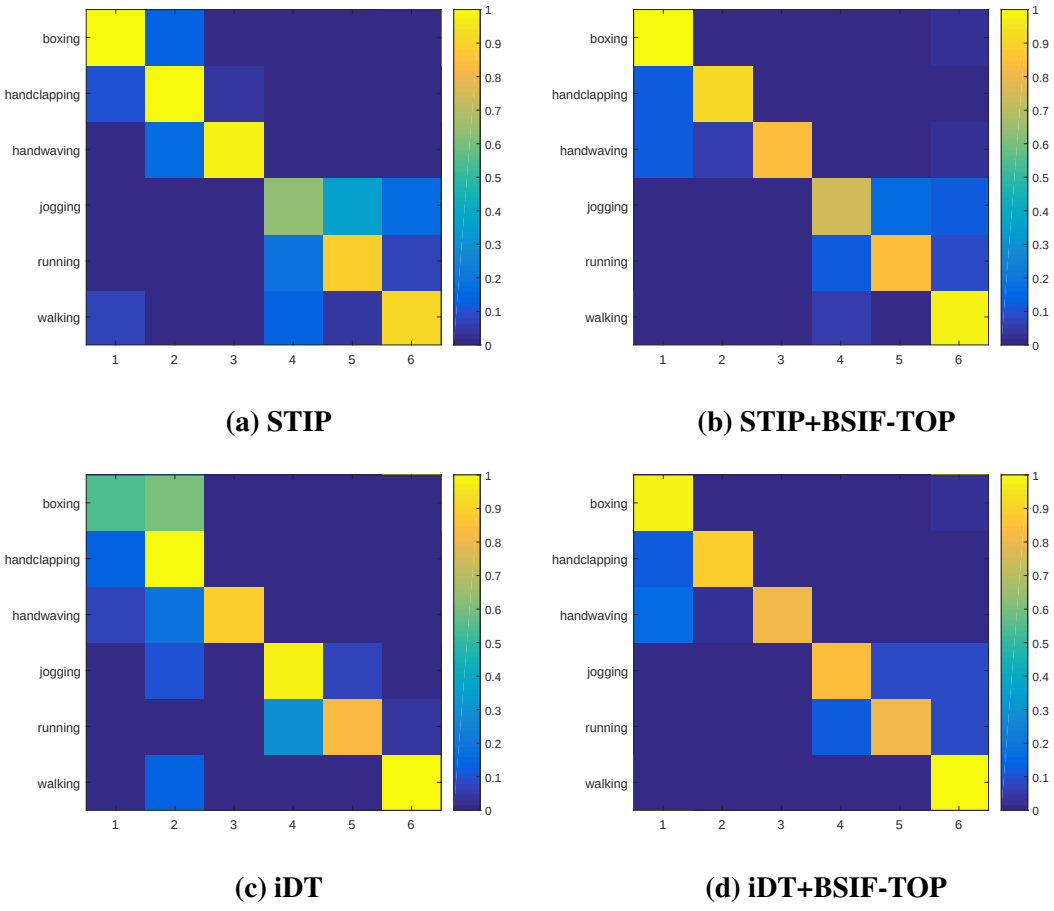


Figure 4.4: Percentage improvement of BSIF-TOP over LBP-TOP and LPQ-TOP, when combined with iDT

The sample confusion matrices for STIP, STIP+BSIF-TOP, iDT and iDT+BSIF-TOP for SD_3 videos are shown in Figure 4.5a, 4.5b, 4.5c and 4.5d. For both STIP and iDT features, it is clear to see that use of textural features helps certain action classes such as *walking* and *jogging* to improve their accuracy by more than ≈20-40%. These two types of videos are very similar to each other, i.e. both has leg movements but



(a) STIP **(b) STIP+BSIF-TOP**

(c) iDT **(d) iDT+BSIF-TOP**

Figure 4.5: Confusion matrix of KTH-SD_3 videos; (a,c) STIP and iDT, (b,d) effects of utilizing textural features (BSIF-TOP) on STIP and iDT. (Best viewed in color)

only speed, and it is difficult to differentiate between them if image structure gets distorted. The feature detectors detects many wrong or non-action relevant points that mostly from background due to this problem. Eventually, this hampers the discriminative capacity of the features extracted which lead us towards this poor result. However, in that case, textural features extracted from those distorted videos still holds quite distinguishable cues that helps to improve the performance. The videos in *boxing*, *handclapping* and *handwaving* may look different from each other, but they have one thing in common – only the video parts belongs to the hand has movements, other parts has no visible movements. Whenever, image structure gets distorted, many of these videos acts ambiguous to each other (highly noticeable at (a) and (c)). In this scenario, texture helps to reduce the ambiguity between them, specially to *boxing* class since it

is a different in nature than the two others. The *handwaving* videos is still ambiguous with both *boxing* and *handclapping* videos, and the *handclapping* videos are ambiguous to *boxing* videos. This is less apparent if image structures becomes good. Overall, the use of textural features on these videos helps to improve the performance by a very good margin, which about more than ≈50%.

Experiments on compressed videos of UCF-11 dataset (YoutTube-LQ): In this section, the same experiments were repeated on compressed videos of UCF-11 dataset (refereed as YoutTube-LQ) to demonstrate the effectiveness of textural features on both STIP and iDT features. It is observed that after applying compression, both STIP and iDT features struggle to maintain their original performances, i.e. 71.94%, for STIP and 81.58% for iDT. From the results reported in Table 4.3, it is clear that methods that use additional textural features demonstrate significant improvement, as compared to baseline methods. However, the methods that use iDT features outperforms STIP based methods. Among the textural features, BSIF-TOP outperform the LBP-TOP and LPQ-TOP features.

After applying compression, the performance of baseline features became lower than the original videos due to the deterioration of video frame. By comparing the performance loss, iDT features dropped more than STIP features (iDT - ≈7.5% and STIP - ≈4.5%). However, after applying textures, the performance of both STIP and iDT features improved significantly. Among the textural features, BSIF-TOP is the most promising choice as it offers the highest performance improvement in comparison to LBP-TOP and LPQ-TOP. However, it is also noticeable that for STIP, LBP-TOP performs slightly better than LPQ-TOP, but this is clearly opposite in the case of iDT.

Figure 4.6 gives a closer look at how BSIF-TOP improves in performance over other textural features. BSIF-TOP performs better than the LBP-TOP and LPQ-TOP features as they combined with both the STIP and iDT features. However, on STIP the percentage of improvement is a bit higher than the iDT features. On STIP features, it is about ≈5.5% better than the LBP-TOP and ≈6.5% better than the LPQ-TOP. And

Table 4.3: Recognition accuracy (%) of various feature combinations on and Youtube-LQ dataset.

Method	YouTube-LQ	Method	YouTube-LQ
STIP (Baseline)	67.57	iDT (Baseline)	74.04
STIP+LBP-TOP	70.69	iDT+LBP-TOP	75.59
STIP+LPQ-TOP	69.13	iDT+LPQ-TOP	76.02
STIP+BSIF-TOP	**76.05**	iDT+BSIF-TOP	**80.45**

on iDT, it is ≈5% and ≈4.5% better than the LBP-TOP and LPQ-TOP respectively.

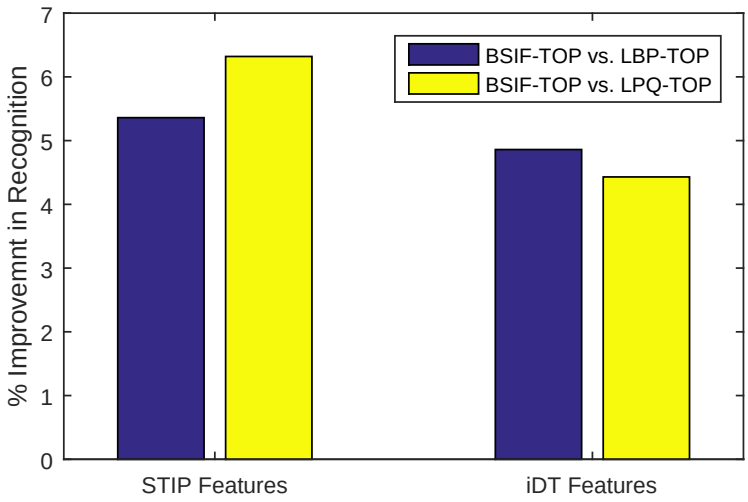
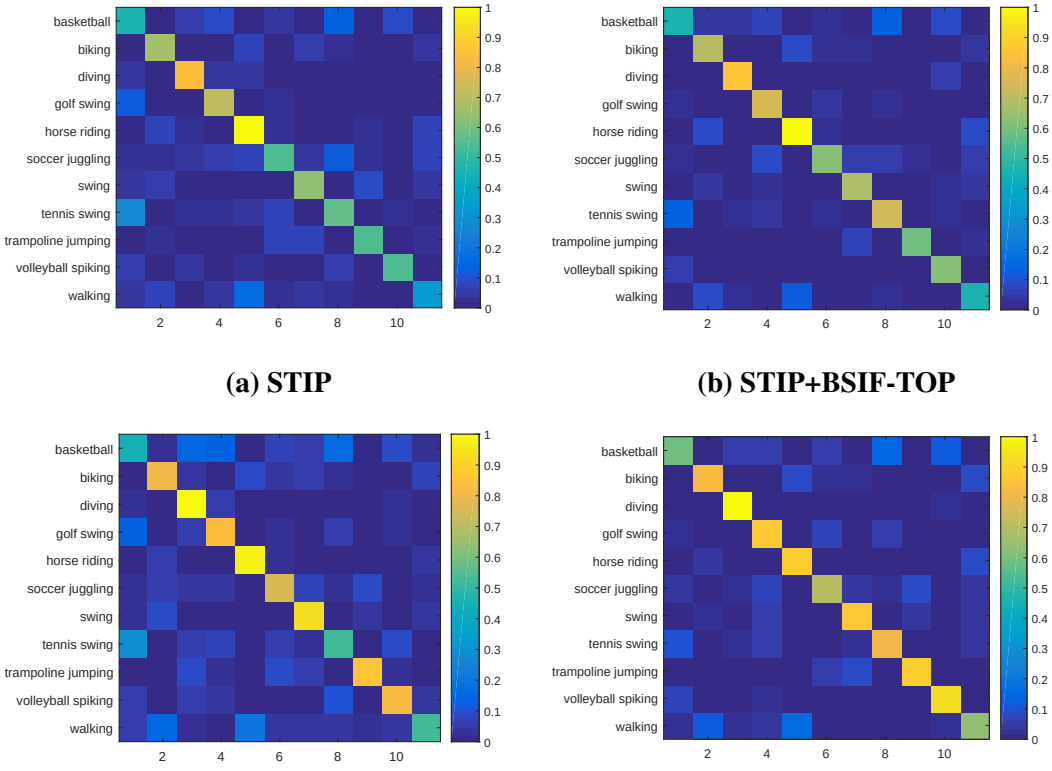


Figure 4.6: Percentage improvement of BSIF-TOP over LBP-TOP and LPQ-TOP, when combined with trajectory based features

Figures 4.7a, 4.7b, 4.7c and 4.7d shows the confusion matrices for STIP, STIP+BSIF-TOP, iDT and iDT+BSIF-TOP, obtained with the *YouTube-LQ* dataset. Similar to KTH videos, it is noticeable that actions that share similar characteristics or movements or objectives such as *basketball* and *tennis swing* are both gaming action and played in stadium becomes ambiguous to each other whenever image structure gets distorted. It is clear that on that kind scenario, textural features have a very good influence. Many action classes have been improved such as *golf swing*, *soccer juggling*, *swing*, *tennis swing*, *trampoline jumping* and *volleyball spiking*. It is interesting to mention that between STIP and iDT, iDT features performed slightly better on actions with com-

plex scenes such as *volleyball spiking* than the STIP features, when combined with BSIF-TOP features.



(a) STIP **(b) STIP+BSIF-TOP**

(c) iDT **(d) iDT+BSIF-TOP**

Figure 4.7: Confusion table of YouTube-LQ; (a,c) STIP and iDT features, (b,d) effects of utilizing textural features (BSIF-TOP) on STIP and iDT features.

Experiments on medium and bad quality subsets from HMDB51 dataset (HMDB51-MQ and HMDB51-LQ): In order to demonstrate the effectiveness of textures on STIP and iDT features for larger action classes, previous experiments were further repeated on low quality subsets of HMDB51 dataset. From the results shown in Table 4.4, it clear that in comparison to baselines, methods use textural features shows a significant leap of performance. However, in comparison to STIP features, methods that uses iDT features show better performance. Like previous experiments, BSIF-TOP offers the most promising results and also outperform the others.

On both STIP and iDT features, in comparison to 'Medium', 'Bad' quality

Table 4.4: Recognition accuracy (%) of various feature combinations on HMDB bad and medium quality subsets.

Method	Bad	Medium	Method	Bad	Medium
STIP (Baseline)	21.71	23.68	iDT (Baseline)	23.88	41.43
STIP+LBP-TOP	20.80	24.28	iDT+LBP-TOP	30.34	43.11
STIP+LPQ-TOP	23.89	28.36	iDT+LPQ-TOP	30.96	42.97
STIP+BSIF-TOP	**32.46**	**37.14**	iDT+BSIF-TOP	**37.80**	**45.96**

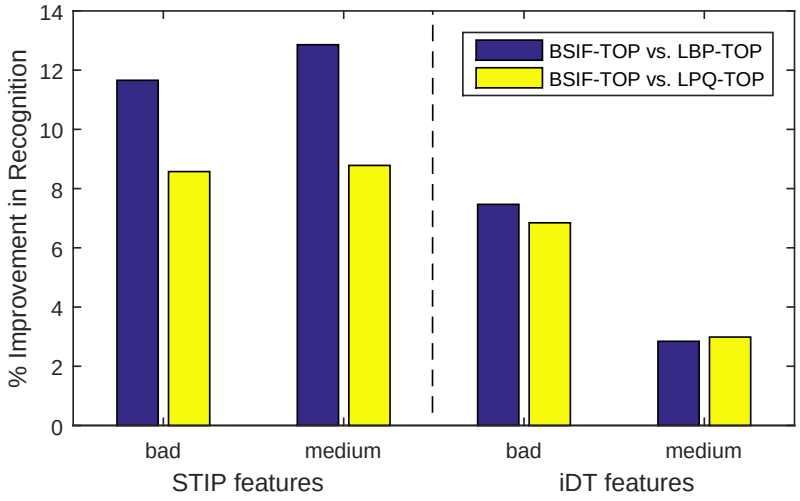
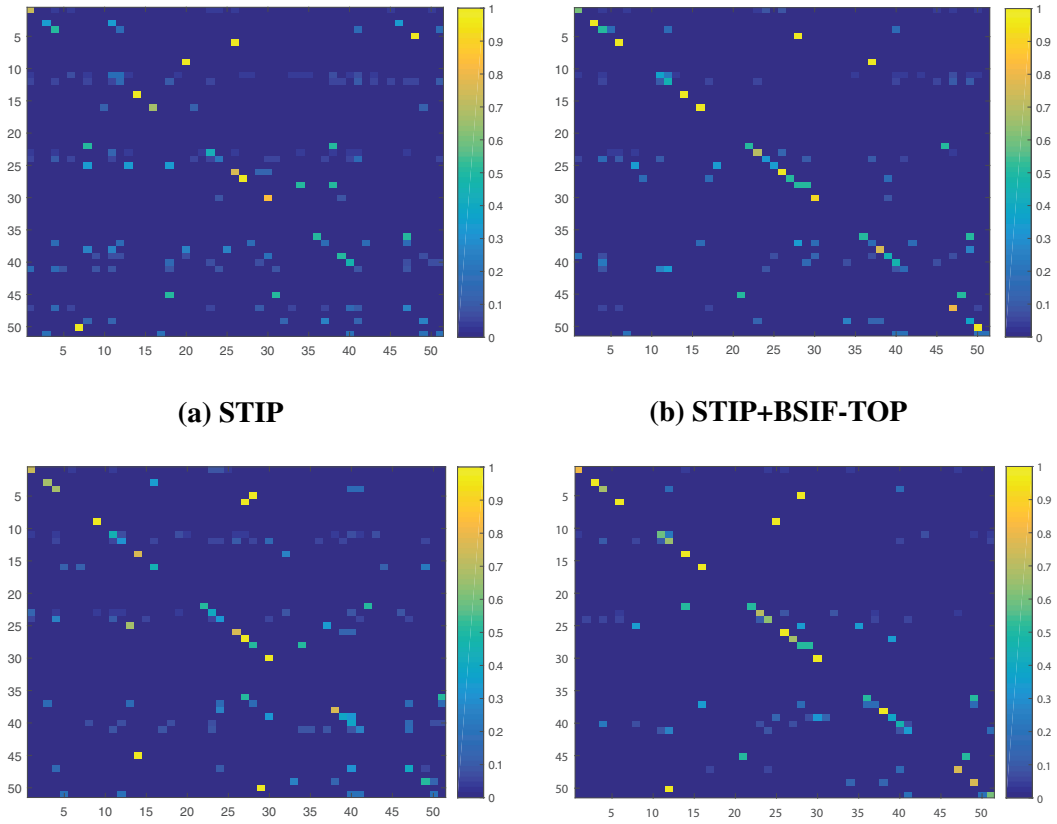


Figure 4.8: Percentage improvement of BSIF-TOP over LBP-TOP and LPQ-TOP, when combined with interest point based features

video yields the lowest performance. The use of textural features on these videos helps to increase their performance by a very good margin. On 'Bad' quality videos, the performance increased by ≈11% and ≈14% when texture is paired with STIP and iDT features respectively. However, for 'Medium' quality videos, performance improvement by textures on iDT is not as great as that on STIP features, but it still offers the highest performance. The BSIF-TOP offers the highest performance among all evaluated textural features for both 'Medium' and 'Bad' quality videos. Between the LBP-TOP and LPQ-TOP, LPQ-TOP shows slightly better performance on 'Bad' quality videos. However, for 'Medium' quality videos, the LBP-TOP perform slightly better than LPQ-TOP when paired with iDT features.

Figure 4.8 gives a closer look at the performance improvement of the BSIF-



(a) STIP **(b) STIP+BSIF-TOP**

(c) iDT **(d) iDT+BSIF-TOP**

Figure 4.9: Confusion matrix obtained from fist split of HMDB dataset; (a,c) interest point and trajectory based shape-motion features, (b,d) effects after after combing BSIF-TOP features. (Best viewed in color).

TOP over the LBP-TOP and LPQ-TOP features. On the 'bad' quality videos, BSIF-TOP is ≈8.5% and ≈6.5% better than LPQ-TOP when paired with STIP and iDT features respectively. On the 'medium 'quality videos, it is ≈12.5% and ≈2.5% better than LBP-TOP when combined with STIP and iDT features respectively.

Figures 4.9a, 4.9b, 4.9c and 4.9d show the confusion matrices for both STIP, STIP+BSIF-TOP and iDT, iDT+BSIF-TOP features, obtained from the first split of HMDB51 dataset. From confusion matrices, it is clear to see that the methods use textures show larger diagonal values which means that it reduces the amount of false positives, hence improving the performance and reducing ambiguity. A total of 15 and 17 action classes improved after applying BSIF-TOP on STIP and iDT feature

respectively. Some action classes that have improved their accuracies are *climb*, *sword*, *draw sword*, *fencing*, *golf* etc.

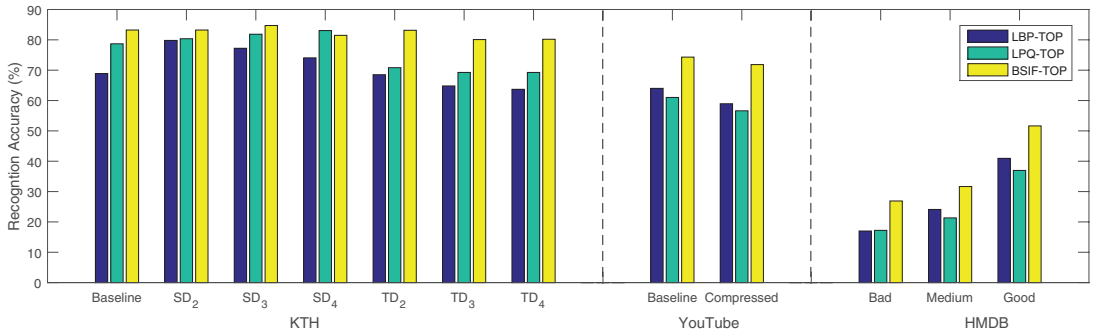
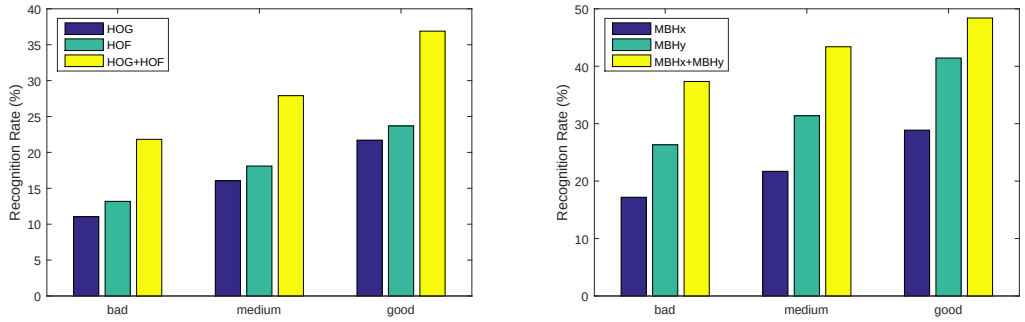


Figure 4.10: Performance of various textural features on KTH and its downsampled versions



(a) STIP features **(b) iDT features**

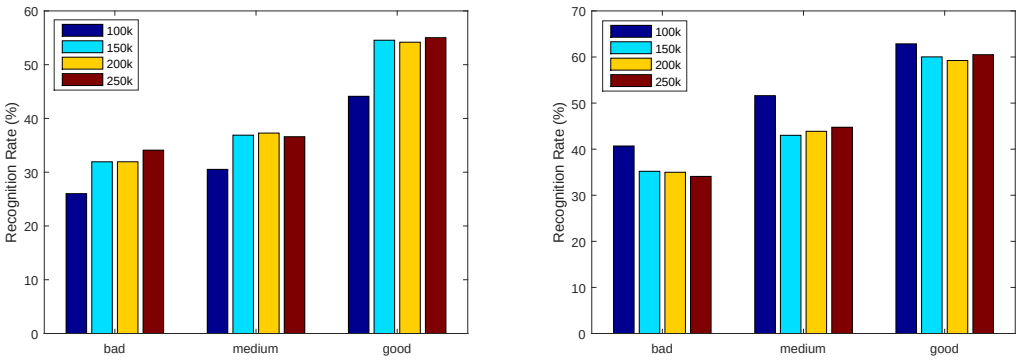
Figure 4.11: Performance comparison of individual and combined use of various shape-motion features on HMDB good, medium and bad quality videos

Analysis on textural features: In this section, an extensive evaluation of various textural features is provided. The analysis could provide further insight into their behavior towards low quality videos. The performance of various textural features across various datasets (baseline and corresponding low quality versions/subsets) is shown in Figure 4.10. From the results, it is clear to see that in comparison with other textural features, BSIF-TOP remain consistently strong across various video quality problems. In certain cases, the BSIF-TOP even outperform the baseline such as in the case of SD_3. Between the LBP-TOP and LPQ-TOP, LPQ-TOP seems to be performing well with downsampled videos and LBP-TOP works well with videos that has a

relatively better spatial and temporal quality. In certain cases, textures, can outperform the shape-motion features. For instance, on HMDB51, BSIF-TOP outperformed the baseline STIP features. However, iDT feature performances are still higher than that of the textures.

Analysis on feature sampling and codebook generation of shape and motion features: Determining the appropriate visual codebook size is an important step for codebook generation. Many authors have analyzed this issue (X. Wang et al., 2013; H. Wang et al., 2009) and suggested to use codebook sizes based on the empirical evaluations over the experimental datasets. Following their suggestions, we run several experiments on various datasets and choose to use k=4000 as codebook size across all datasets. For consistency of experiments, random 100,000 features are chosen from all training samples for building codebook, which gives a decent level of accuracy across various datasets. It is not always expected that sampling of a large feature set will always result in better performance; infact, it increases the computational cost.

Figure 4.12 shows the performance of our two best performing methods on all three HMDB51 subsets, with respect to various feature sampling sizes. On STIP+BSIF-TOP shown in Figure 4.12a, it is clear to see that the recognition accuracy of various HMDB51 videos can be improved with the use of the larger feature sampling size. Most importantly, the 'bad' and 'medium' can obtain better accuracy ($\approx$5%) if larger sampling size is used. However, the scenario for iDT+BSIF-TOP shown in Figure 4.12b is slightly different. The recognition accuracy significantly drops across all subsets, specifically for 'bad' and 'medium' quality subsets with the use of larger sampling sizes. The strategy for codebook generation that we use is to rigidly quantize each feature descriptor by a single basis (or codeword). In cases where features are ambiguously located between two or more codewords, small changes to the feature (due to noise, variance) can lead to different codewords being assigned, thus affecting classification performance. It is worth mentioning that the large sampling of very different type of feature combination, i.e. STIP, performs better than the similar type of feature combination, i.e. iDT.



(a) STIP+BSIF-TOP **(b) iDT+BSIF-TOP**

Figure 4.12: Recognition performance of (a)STIP+BSIF-TOP and (b)iDT+BSIF-TOP approaches on various subsets of HMDB dataset, with respect to various amount of feature descriptor sampling

Analysis on shape and motion features encoding: Many action feature encoding methods have been proposed in literature such as histogram encoding, fisher vectors and sparse coding (X. Wang et al., 2013), and we choose to use the histogram encoding (BoVW) which is popularly used by many action recognition methods (H. Wang et al., 2009, 2011). Though some action recognition methods (H. Wang et al., 2011; H. Wang & Schmid, 2013) show that fisher vector (FV) is superior to the BoVW encoding, but for low quality videos (except videos from HMDB51), we found that the BoVW is more effective than FV. Our best methods (STIP+BSIF-TOP and iDT+BSIF-TOP) achieve higher accuracy if we use BoVW encoding instead of fisher vector encoding, as shown in Table 4.5. It is interesting to mention that the FV does not always hold advantage over BoVW, if spatial resolution is reduced or compressed. However, it is still better than BoVW in case of complex scenes (i.e., HMDB51), and when temporal sampling rate deteriorates.

Analysis on computational complexities: A comparison between the speed of various feature descriptors (including both their feature detection and quantization time) is given in Table 4.6. The comparison was performed on a sample video from *bike riding* action class of HMDB51 dataset that consist of a total of 246 image frames (240×320 pixel resolution) at 30 frames per second. The estimation of run-time was performed using MATLAB 2015a on a Intel Core i7 3.60 GHz machine with 24GB

Table 4.5: Recognition accuracy (%) of various datasets with STIP+BSIF-TOP and iDT+BSIF-TOP methods using bag-of-visual-words (BoVW) and fisher vector (FV) encoding.

	STIP+BSIF-TOP		iDT+BSIF-TOP	
Datasets	BoVW	FV	BoVW	FV
KTH-SD_2	88.80	**89.26**	**93.89**	92.87
KTH-SD_3	**85.28**	83.15	**88.33**	87.78
KTH-SD_4	**81.67**	80.19	**82.41**	81.02
KTH-TD_2	88.70	**89.91**	**95.09**	94.44
KTH-TD_3	86.11	**87.78**	92.22	**92.59**
KTH-TD_4	**84.54**	82.96	90.00	**90.28**
Youtube-LQ	**76.05**	75.04	**80.45**	78.13
HMDB-BQ	32.46	**33.06**	37.80	**40.69**
HMDB-MQ	37.14	**38.51**	45.96	**51.62**

RAM. Table 4.6 shows the computational cost of various feature descriptors per image frame. Among the shape-motion descriptors, the STIP descriptors takes ≈0.047 seconds faster time than iDT which is expected, since MBH relies on feature tracking and warped flow estimation. Between the textural features used, the LPQ-TOP and BSIF-TOP are the most efficient methods (both much faster than the computation of the shape-motion features), and yet they are able to contribute significantly to the recognition accuracy.

Table 4.6: Computational cost (detection/calculation + description) of various feature descriptors per image frame

	STIP	iDT	LBP-TOP	LPQ-TOP	BSIF-TOP
Time (in sec.)	0.156	0.203	1.230	0.041	0.051

4.5 Summary

In this chapter, a framework that exploits textural features to improve the performance of action recognition in low quality videos is presented. In comparison with current methods that are mainly relying on shape and motion features, the complementary use of textural features is a novel proposition that improves the recognition

performance by a good margin. Though there is high relevancy of low quality videos in real-world application, but to best of our knowledge, there are no systematic works that investigate this specific problem of video quality. By our scope, we considered videos that are poorly sampled in spatial or temporal domains, compressed, and are adversely affected by motion blur and camera motions. This work draws some interesting observations as to how low quality videos can particularly benefit from textural information, considering that most new approaches involve only shape and motion information as their choice of space-time features.

In the future, we intend to explore other textural features that are potentially more robust towards the deterioration of video quality and are more efficient. Textural features that are denser or richer in description are also potential directions following this work. Also, it is worth exploring visual features, i.e. shape, motion or texture that offers action representation with high discriminative capacity in low quality condition.

CHAPTER 5

SPATIO-TEMPORAL MID LEVEL FEATURE BANK FOR ACTIVITY RECOGNITION IN LOW QUALITY VIDEO

Many methods for activity recognition have been proposed in recent literature (H. Wang et al., 2009; X. Wang et al., 2013; H. Wang & Schmid, 2013). Most of them are *handcrafted* methods and they represent video in the form of shape and motion features. The feature representation scheme of these methods is designed for high quality videos, and is not appropriate for low quality videos. Current methods are increasingly ineffective when the quality of the video deteriorates in both spatial and temporal domain. This problem can be alleviated by using complementary global textural features (in Chapter 4). However, the holistic nature of these global descriptors often produces indiscriminative features since illumination variations or unrelated motion from videos with complex backgrounds can be erroneously regarded as textures.

In this chapter, a new spatio-temporal mid-level (STEM) feature bank for recognizing actions in low quality videos is proposed. STEM is designed to combine the benefits of local explicit patterns surrounding interest points, and global statistical patterns. Specifically, we first detect features at the local and global levels, which we call *streams*; for each stream, state-of-the-art spatio-temporal approaches (Laptev, 2005; Kannala & Rahtu, 2012) are used to describe their respective patterns. More significantly, the textural features are compactly extracted from 3D salient patches. To build the STEM feature bank from a trio of features, shape-motion descriptors soft-quantized by Fisher Vector encoding, and discriminative textural descriptors are integrated.

In section 5.1, related methods and motivations behind the design of STEM is discussed. Then, in section 5.2, the proposed Spatio-temporal Mid Level Feature Bank (STEM) framework is discussed. Various components and feature representation methods are also discussed in the same section. In section 5.3, the evaluation frame-

work and the parameters used in STEM are discussed. In section 5.4, experimental results of STEM and further analysis are shown. Finally, section 5.5, summarizes this chapter.

5.1 Related Work and Motivations

The literature (Aggarwal & Ryoo, 2011; S. Vishwakarma & Agrawal, 2013; H. Xu et al., 2015) is packed with various types of activity recognition methods. Among handcrafted methods, shape and motion feature feature based methods become very popular for their excellent results. A detailed review of recent methods is presented in Section 4.1 of Chapter 4. Generally, the extraction of these features consists of two distinct steps: *detection* and *description*. In *detection* step, important points or salient regions are extracted from the video, and the patterns of the extracted points/regions are then described in *description* step. Among typical detectors include space-time interest points (Laptev, 2005), cuboids (Dollár et al., 2005), dense sampling (H. Wang et al., 2009), dense trajectories (H. Wang et al., 2011), and improved dense trajectories (H. Wang & Schmid, 2013). The HOG, HOF and MBH features appeared most prominently in recent state-of-the-art approaches (Laptev, 2005; H. Wang et al., 2009; H. Wang & Schmid, 2013) owing to their effectiveness in characterizing structural and dynamic properties of human activities. However, their reliance on localized feature regions may render them ineffective when discriminating between actions in low video quality.

The use of *textural* features is less common in action recognition; among proposed representations include LBP-TOP (Kellokumpu et al., 2008a) and Extended LBP-TOP (Mattivi & Shao, 2009). These methods use the notion of three orthogonal planes (TOP) to extend static image-based textures to spatio-temporal dynamic textures. More recently, Kannala and Rahtu (2012) proposed binarized statistical image features (BSIF) which have shown tremendous potential compared to its predecessors. While these methods were able to show effective results across different action datasets, our evaluation in Chapter 4 has shown that local STIP features (Laptev, 2005) become increasingly ineffective when video quality deteriorate spatially and tempo-

rally. It was shown that this problem can be alleviated by introducing complementary robust global textural features. However, the holistic nature of these descriptors often produces indiscriminate features since illumination variations or unrelated motion from videos with complex background can be erroneously regarded as textures.

Salient and discriminative feature section remains a challenging issue for activity recognition community. Many efforts concerning salient feature formation have been reported in literature. Among various methods, Q. Wu et al. (2013) used saliency maps to prune irrelevant spatio-temporal interest points (STIPs) – STIPs that mostly belongs to the background. Yi and Lin (2013) and X. Chen et al. (2015) used saliency maps to prune the activity-irrelevant trajectories. All of these methods appear to suggest that the use of salient maps helps to improve the feature performances. It also increases the discriminative capacity of the features. However, if the detected feature points are less, use of saliency maps may reduce the recognition performance (Q. Wu et al., 2013).

Mid-level or body parts features are also used by many activity recognition methods, but they are not as popular as low-level features. Among recently proposed methods, A. Jain et al. (2013) used discriminative mid-level patches and L. Wang et al. (2013) used spatio-temporal mid-level body parts named 'motionlets'. In both these works, salient patches are first detected, then discriminative patches are selected using clustering or data mining techniques. Selected patches are then used for representation of activities. However, detection of discriminative patches is quite challenging when videos with complex background and low quality videos are considered, and it may drastically affect the recognition performance.

The concept of feature bank have been proposed by activity recognition community in recent years. A feature bank is a collective repository of various features (from high level to low level) extracted from a video. Among the proposed methods, that adopted the concept of feature bank is Action bank (Sadanand & Corso, 2012), which represents each video by a collective form of many detectors that produces a

correlation volume. The basic component of action bank is a template based detector that detects actions. Since it is based on the template, action from varying scales, viewpoints and tempos were detected. Action bank represents videos in a form of high level features. The method achieved a very high accuracy (98.2%) for the KTH dataset (Schüldt et al., 2004). However, action bank does not work well on videos with complex scenes such as HMDB51 dataset (26.9%). Also, formation of the collection of detectors is computationally costly.

Motivated by the analysis above, a new spatio-temporal mid-level (STEM) feature bank is proposed in this work for recognizing actions in low quality videos. STEM is designed to combine the benefits of local explicit patterns surrounding interest points, and global statistical patterns. Specifically, at first, features at the local and global streams are detected, then for each stream, state-of-the-art spatio-temporal approaches (space-time interest points (Laptev et al., 2008) or improved dense trajectories (H. Wang & Schmid, 2013), and binarized statistical image features (Kannala & Rahtu, 2012)) are used to describe its respective patterns. More significantly, the textural features are compactly extracted from 3D salient patches.

5.2 Proposed Spatio-temporal Mid Level Feature Bank

A graphical overview of our proposed feature bank is illustrated in Figure 5.1. An input video undergoes a series of steps, in two separate streams: *local* and *global*. In the local stream, spatio-temporal interest points or dense trajectories are first detected, and represented with shape-motion descriptors. The global stream involves the extraction of spatio-temporal textural features, which are discriminately selected based on visual saliency. To obtain a compact mid-level representation, the local shape-motion features are encoded by Fisher Vectors (FV), while the global textural features are built based on the regions defined by the 3D saliency mask. The concatenation of both feature sets produces the spatio-temporal mid-level (STEM) feature bank, which is subsequently used to represent the video for classification. We present the shape-motion features and salient textural features in sections 5.2.1 and 5.2.2 respectively.

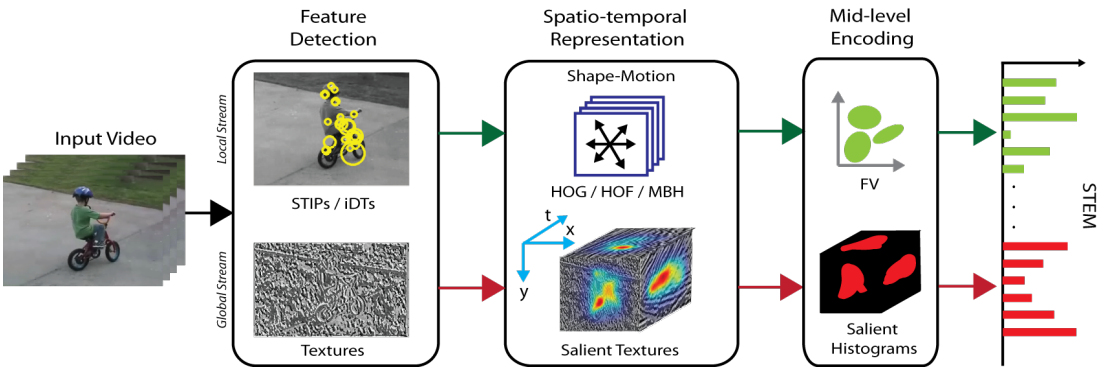


Figure 5.1: Illustration of the proposed STEM feature bank

5.2.1 Shape-Motion Features

Two types of spatio-temporal features are extracted from video namely, (1) interest point based features and (2) trajectory based features. To extract interest point based features from video, we employed Harris3D detector (Laptev, 2005) to detect spatio-temporal interest points (STIPs). To describe patterns across STIPs, HOG and HOF feature descriptors (Laptev et al., 2008) are used. For estimation of trajectory based features, improved dense trajectories (iDT) (H. Wang & Schmid, 2013) is used, and description of motion pattern on the trajectories uses the MBH feature descriptor. The detailed description of detectors and descriptors can be found in Section 4.2.1 of Chapter 4.

5.2.2 Salient Textural Features

This section describes a new spatio-temporal textural representation, called salient textures. At first, the estimation process of textures is given and then its spatio-temporal representation is discussed. After that, details of the spatio-temporal salient texture estimation are discussed.

Textural feature detection and representation: Binarized statistical image features (BSIF) (Kannala & Rahtu, 2012) is a recently proposed method that efficiently encodes texture information, in a similar vein to earlier methods that produce binary codes such as local binary pattern (LBP) (Ojala et al., 2002) and local phase

quantization (LPQ) (Päivärinta et al., 2011). Inspired by its efficient feature encoding capacity as shown in Chapter 4, BSIF features are used for estimating texture in the STEM framework. The detailed description of BSIF features is given in Section 4.2.2 of Chapter 4.

However, using space-time volume (STV), the BSIF method offers only structural or shape information which does not have high discriminative capacity for recognizing human activities in video. Following the success of three orthogonal planes (TOP) approach in extracting spatio-temporal dynamic textures (G. Zhao & Pietikainen, 2007; Päivärinta et al., 2011) in Chapter 4, this work applies the same TOP approach to extract spatio-temporal textures. The detailed description of TOP approach is given in Section 4.2.2 of Chapter 4. The BSIF-TOP textural feature is of length $3 \cdot 2^n$, where n is number of bit size.

In this work, BSIF-TOP$_{9,12}$ was used as parameter settings, which was chosen empirically, across all video samples. An illustration of TOP method using BSIF features is shown in 5.2.

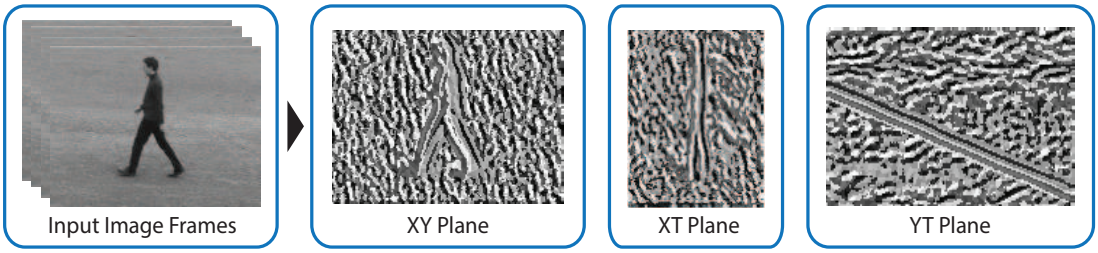


Figure 5.2: A sample image frame and its corresponding BSIF code images in the XY, XT and YT planes

Salient texture formation: In an action video such as *running* or *fencing*, the runner or fencers gets the most visual attention across the video, hence they correspond to the most salient locations in video. As such, we apply saliency to the textural features extracted from each video. Motion or video-based saliency methods are not always effective in constructing equitable saliency maps in presence of large camera

motion and noisy flow fields (Sultani & Saleemi, 2014). In this work, graph-based visual saliency (GBVS) (Harel et al., 2006) is used to capture the salient regions in each frame. This method is computationally simple and is able to model natural fixation based on a variety of features.

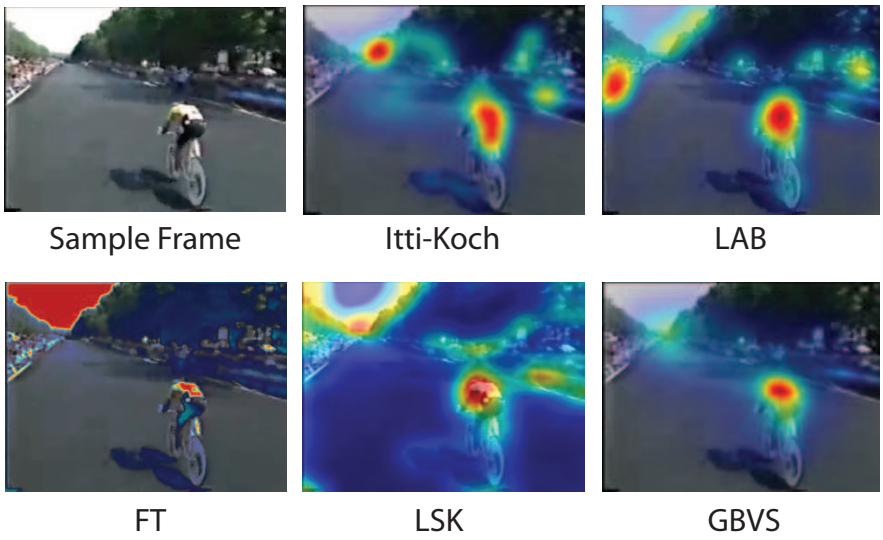


Figure 5.3: A sample image frame from 'biking' activity class of YouTube-LQ dataset (J. Liu et al., 2009) and corresponding feature maps using various saliency methods.

The GVBS is used due to its efficiency in localizing human activities in both simplistic and complex scene structures. Figure 5.3 shows the saliency maps generated by various popular methods, i.e. *Itti-Koch* saliency model (Itti et al., 1998), Signature driven *LAB* (Hou et al., 2012), Frequency tuned (*FT*) saliency (Achanta et al., 2009), and local steering kernel (*LSK*) (Seo & Milanfar, 2009); on a complex scene structure (complex in terms of shooting environment, i.e. shadow, occlusion, low resolution etc.). Many of these methods are also used in a recent work named STAP (Nguyen et al., 2015) that uses multiple saliency maps to pool video features with the visual attention. From the figure, it is noticeable that, except for GVBS, most of the methods predict many actions-irrelevant parts that mostly belongs to background as a salient part of the image. The output of the LAB method is close to that of GVBS, but it also predicts many irrelevant image parts which may cause feature descriptors to lose some discriminativeness. As such, we opt for the GBVS due to its simplicity in computation

and good saliency prediction in video frames.

Firstly, a set of three feature maps is calculated: A *contrast* map computed using luminance variance in a local neighborhood of size one-tenth of the image width, four *orientation* maps computed using Gabor filters (at orientations 0°, 45°, 90°, 135°), and a *flicker* map by simple frame differencing. Next, a fully connected directed graph $\mathscr{G}$ is constructed to extract activation maps, where the weight of edge between two nodes corresponding to pixels at location(x,y) and (m,n) can be expressed as:

$$w^{act}_{x,y,m,n} = |M_{x,y} - M_{m,n}|\, e^{\left(-\frac{(x-m)^2+(y-n)^2}{2\sigma_1^2}\right)} \tag{5.1}$$

where σ_1 is a free parameter. The calculated graph with normalized weights is used to construct a Markovian chain (Harel et al., 2006) where achieving equilibrium distribution results in an activation map $A_{m,n}$. To normalize the activation map, another graph is constructed with each coordinate from $A_{m,n}$ representing the nodes, and the weights are defined as,

$$w^{norm}_{x,y,m,n} = A_{m,n}\, e^{\left(-\frac{(x-m)^2+(y-n)^2}{2\sigma_2^2}\right)} \tag{5.2}$$

When equilibrium distribution is achieved, mass will be concentrated at nodes with high activation, thus forming the final saliency map, $S_{x,y}$.

Finally, for each frame the saliency map $S_{x,y}$ is converted into a binary saliency mask $Z_{x,y}$ by utilizing Otsu's method (Otsu, 1975). This method was chosen because of its capability of locating an optimal threshold that optimizes the foreground and background pixel distributions. This yields a 3D salient mask, $Z(p)$ with p a pixel at location (x,y,t). Applying saliency to Eq. 4.9, computation of the j-th bin for the new salient BSIF histogram is redefined as

$$\bar{h}_j^{plane} = \sum_{p \in plane} \mathscr{I}\{\{b_i(p) = j\} \cap \{Z(p) = 1\}\} \tag{5.3}$$

5.3 Evaluation Setup

This section defines the evaluation parameters and methods used. Following the works of X. Wang et al. (2013) and Perronnin, Sánchez, and Mensink (2010), $k = 256$ is used for FV encoding, applying power and ℓ_2–normalization (a detailed description of FV and normalizations are given in Section A.2 and A.3 of Appendix 1). The free parameters in Eq. 5.1 and 5.2 (σ_1, σ_2) are fixed to 0.15 and 0.06 of the map width respectively, following the author's implementation (Harel et al., 2006). For classification, the action labels are determined by employing a multi-class support vector machine (SVM) with homogeneous χ^2-kernel (Vedaldi & Zisserman, 2012), adopting a one-versus-all strategy.

5.4 Experimental Results and Analysis

This section presents the experimental results of the proposed STEM feature bank on low quality versions – the six KTH downsampled versions and YouTube-LQ, and HMDB51 subsets – HMDB51-MQ and HMDB51-BQ. Tables 5.1, 5.2, 5.3 and 5.4 show all the results. The comparisons with other recent approaches are also provided.

Results on KTH downsampled dataset: Here, KTH dataset was chosen to perform this extensive downsampled experiment as it is lightweight and a widely used benchmark among most public datasets. The results for STIP based STEM features are shown in Table 5.1, while iDT based STEM features are shown in Table 5.2. It is observed that as the video quality deteriorates (particularly the spatial resolution), most methods struggle to maintain their original performances, i.e. 92.13%, for X. Wang et al. (2013), 91.8% for H. Wang et al. (2009) and 95.46% for H. Wang and Schmid (2013); except the iDT based STEM features on SD_2. The proposed STIP based STEM features outperforms other feature approaches across all modes except in SD_2, while iDT based STEM feature outperform other approaches only in SD_3 and SD_4 modes. However, the iDT feature based STEM outperforms STIP based STEM across all video modes. This shows that iDT based STEM features are more robust towards poorer video quality; the improvement is most obvious when dealing with spatial resolution.

Table 5.1: Recognition accuracy (%) of STEM using STIP features on various feature approaches on the spatially and temporally downsampled KTH low quality versions.

Method	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
STIP[1](H. Wang et al., 2009)	88.24	81.11	73.89	87.04	82.87	82.41
STIP[2] (X. Wang et al., 2013)	89.63	82.31	78.98	89.35	86.11	83.89
Deep Obj. (Rahman et al., 2016)	89.81	82.41	81.94	87.50	85.65	82.87
STIP+LBP-TOP	**89.81**	81.48	78.70	89.35	86.11	84.72
STIP+LPQ-TOP	87.41	80.19	76.30	87.41	81.85	79.81
STIP+BSIF-TOP	89.35	82.87	79.72	89.63	87.41	84.63
STEM (based on STIP)	88.52	**83.98**	**83.15**	**90.00**	**88.06**	**85.09**

Table 5.2: Recognition accuracy (%) of STEM using iDT features on various feature approaches on the spatially downsampled and temporally downsampled KTH low quality versions.

Method	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
iDT[1] (Peng et al., 2014)	92.59	78.80	61.85	95.19	91.57	89.54
iDT[2](H. Wang & Schmid, 2013)	94.07	79.91	64.17	94.63	92.50	89.17
Deep Obj. (Rahman et al., 2016)	94.91	90.74	84.26	94.91	92.13	90.74
iDT+LBP-TOP	**94.26**	80.00	69.91	**94.63**	92.59	89.91
iDT+LPQ-TOP	94.07	80.00	78.80	94.63	92.59	89.63
iDT+BSIF-TOP	92.87	87.78	81.02	94.44	92.59	**90.28**
STEM (based on iDT)	93.24	**88.98**	**83.89**	94.54	**92.59**	89.81

STIP based STEM outperform joint feature utilization methods proposed in Chapter 4 across all downsampled modes except SD_2, where STIP+LBP-TOP perform slightly better ($\approx 0.29\%$). iDT based STEM also outperforms the joint feature utilization methods, but it only performs better in SD_3, SD_4 and TD_3 modes. Like STIP+LBP-TOP, iDT+LBP-TOP also performs better than the iDT based STEM in SD_2 mode. This suggests that, LBP-TOP features perform better with high spatial resolution. In case of temporally downsampled videos, iDT based STEM performs slightly lower than the joint feature utilization methods. In TD_2, the performance of iDT+LBP-TOP and iDT+LPQ-TOP are similar (94.63%), and both slightly outper-

[1]use Bag-of-Words (BoW) encoding

[2]use FV encoding

form the STEM. However, their performance is still lower than the method used by Peng et al. (2014). Similarly, in TD_4, iDT+BSIF-TOP perform better than the STEM. The results in TD_4 mode suggests that when videos are temporally downsampled at an extreme level, the salient BSIF-TOP does not perform well with iDT features.

For a better perspective, the STEM approach garnered 94.35% using STIP features on the original KTH data, which is only marginally better than the other methods and 95.46% using iDT features, which is similar to H. Wang and Schmid (2013). It can be observed that the discriminative nature of the salient textures in STIP based STEM plays a significant role in obtaining better performance compared to its non-salient counterpart, especially in SD_4 and TD_4 videos. Similarly, for iDT based STEM, use of saliency improves the performance compared to its non-salient counterpart except TD_4. In extreme temporal downsampling situation such as TD_4, where many important frames are removed, pruning of action relevant parts such as backgrounds using saliency hampers the action modeling capacity.

It is interesting to mention that the performance of STEM is better than the combination of STIP and Deep Object Features extracted from ImageNet-pretrained CNN model (Rahman et al., 2016) in almost every downsampled case except for SD_2, where it performed as well as the STIP+LBP-TOP features. However, this is not the same for the combination of iDT and Deep Object Features except for the case TD_3 where its performance is quite close to that of STEM.

Results on UCF-11 compressed videos: To assess the effectiveness of the proposed STEM on compressed videos, the same experiments were repeated on the YouTube-LQ dataset. The performance of STEM on the original YouTube dataset was also investigated, and the drop in performance after applying compression is $\approx$2.81% for STIP based STEM features, and $\approx$6.88% for iDT based STEM features. Between both STEM, the one build on trajectories greatly hampered due to the use of compression in the spatial domain. From the results shown in Table 5.3 and 5.4, it is clear that both the STIP and iDT based STEM outperform joint feature utilization methods

and related state-of-the-arts. However, iDT based STEM outperform the STIP based STEM (by ≈2.03%). This clearly justifies the robustness of both STEMs on recognizing human activities from compressed video.

Among various joint utilization methods, the performance of STIP+BSIF-TOP and iDT+BSF-TOP is close to the performance STEM. The difference is about ≈2.45% for STIP based STEM and ≈1.39% for iDT based STEM. Other methods, including the ones comprise of LBP-TOP and LPQ-TOP features perform less than that. The improvement of performance by iDT based STEM is not as high as STIP based STEM, but it marginally improves the performance.

Table 5.3: Recognition accuracy (%) of STEM using STIP based features on various feature approaches on the YouTube-LQ dataset.

Method	YouTube-LQ
STIP[1] (H. Wang et al., 2009)	67.57
STIP[2] (X. Wang et al., 2013)	70.27
Deep Obj. (Rahman et al., 2016)	86.36
STIP+LBP-TOP	70.99
STIP+LPQ-TOP	71.65
STIP+BSIF-TOP	75.04
STEM (based on STIP)	**77.49**

Table 5.4: Recognition accuracy (%) of STEM using iDT based features on various feature approaches on the YouTube-LQ dataset.

Method	YouTube-LQ
iDT[1] (Peng et al., 2014)	74.04
iDT[2] (H. Wang & Schmid, 2013)	67.10
Deep Obj. (Rahman et al., 2016)	86.82
iDT+LBP-TOP	68.57
iDT+LPQ-TOP	70.59
iDT+BSIF-TOP	78.13
STEM (based on iDT)	**79.52**

Along with compressed videos, STEM also comparatively achieves higher per-

Table 5.5: Recognition accuracy (%) of STEM on UCF-11 dataset.

Method	UCF-11
HOG/HOF[1] (J. Liu et al., 2009)	71.2
DT[1] (H. Wang et al., 2011)	84.2
iDT[2] (only MBH) (H. Wang & Schmid, 2013)	83.6
STEM (based on STIP)	**80.3**
STEM (based on iDT)	**86.4**

formance than other recent methods such as J. Liu et al. (2009), H. Wang et al. (2011) and H. Wang and Schmid (2013) on the original UCF-11 dataset. Table 5.5 show a comparison between these methods and STEMs. The STIP based STEM outperform baseline method (J. Liu et al., 2009) by ≈9.1%, and iDT based STEM outperform H. Wang et al. (2011) and H. Wang and Schmid (2013) by ≈2.2% and ≈2.8% respectively.

The salient textural features contribute greatly in improving the recognition performance, in the case of STIP features. It improves the performance of BSIF-TOP features by ≈2.45%. The improvement in iDT features is not so great as STIP (≈1.39%), although it marginally improves the overall performance. This suggests that the idea of 3D salient textures helps to improve the activity recognition performance in compressed video domain.

It is worth mentioning that Deep Object Features has a great impact when combined with both STIP and iDT features. It greatly improves the performance of both types of features by about ≈9% for STIP and ≈7% for iDT. This may further improve if we use a better CNN model that extracts object-level features more robustly.

Results on HMDB51 subsets: In order to evaluate the effectiveness of the proposed method on a larger number of classes, STEM is also tested on the HMDB51 low quality subsets. Results in Table 5.6 and 5.7 show the superiority of the STEM over other approaches. It can be observed that for both low quality subsets, STEM outperforms other related and joint feature utilization methods. Between the STEMs,

in both subsets, the iDT based STEM outperform STIP based STEM by a great margin (≈6.84% for HMDB51-BQ and ≈12.85% for HMDB51-MQ). Moreover, compared to the baselines (Kuehne et al. (2011) and H. Wang and Schmid (2013)), the improvement of performance by both STEMs in bad quality subset (≈16.38% for STIP and ≈9.94% for iDT based STEM) is higher than the medium quality subset (≈9.94% for STIP and ≈5.44% for iDT based STEM). This shows that proposed STEMs are robust for activity recognition in low video quality.

In comparison with the joint feature utilization methods, the performance of STIP based STEM is very close to the STIP+BSIF-TOP and iDT based STEM is very close to the iDT+BSIF-TOP. The LBP-TOP combined methods only perform better with STIP features, while LPQ-TOP combined methods perform better with iDT features. However, the baseline method by H. Wang and Schmid (2013) which is comprised of iDT features performs better than the LBP-TOP and LPQ-TOP combined methods. Also, both STIP+BSIF-TOP and iDT+BSIF-TOP performs higher, among other joint feature utilization methods.

Table 5.6: Recognition accuracy (%) of STEM using STIP features on various feature approaches on the HMDB51 low quality subsets

Method	BQ	MQ
STIP[1] (Kuehne et al., 2011)	17.40	22.77
STIP[2] (X. Wang et al., 2013)	26.02	30.53
Deep Obj. (Rahman et al., 2016)	33.74	40.55
STIP+LBP-TOP	28.49	35.24
STIP+LPQ-TOP	25.02	30.75
STIP+BSIF-TOP	33.78	38.76
STEM (based on STIP)	**34.08**	**38.94**

The salient textural features which are produced on the global stream clearly help to improve the recognition on both BQ ('bad') and MQ ('medium') subsets. Interestingly, the use of salient textures in STEM could only marginally surpass that of non-salient textures. This is likely due to the complexity of background scenes in this

Table 5.7: Recognition accuracy (%) of STEM using iDT features on various feature approaches on the HMDB51 low quality subsets

Method	BQ	MQ
iDT[1] (Peng et al., 2014)	28.87	41.43
iDT[2] (H. Wang & Schmid, 2013)	30.98	46.35
Deep Obj. (Rahman et al., 2016)	42.01	53.37
iDT+LBP-TOP	30.57	45.43
iDT+LPQ-TOP	30.76	45.96
iDT+BSIF-TOP	40.69	51.62
STEM (based on iDT)	**40.92**	**51.79**

dataset, which made it difficult to obtain good salient patches.

It is highly noticeable that for STIP based approaches, the proposed STEM performs better than adding Deep Object features as proposed in Rahman et al. (2016) in case of HMDB51-BQ. This justifies that STEM is robust to poor quality videos. However, this is not true for iDT based features, where adding Deep Object features yields a better performance than STEM ($\approx$1%). For HMDB51-MQ, the result of combining the baseline features with Deep Object features also comes very close to that of the STEM ($\approx$1-2%) features. This is likely because of better image resolution. Deep Object features are able to provide better discrimination ability when the quality of video improves.

Multi-scale salient texture: To further enrich the statistical information encoded in our salient texture descriptor, texture formation was further extended to a multi-scale variety by employing a number of filters of different sizes, specifically $l = \{3, 9, 15\}$. This is able to increase the accuracy of STEM on all used datasets by $\sim$ 1-2% but at the expense of higher computational load. This direction can be further explored in future work.

5.5 Summary

In this chapter, a new spatio-temporal mid-level (STEM) feature bank that integrates the advantages of local explicit patterns, and global salient statistical patches is presented. The idea of pruning activity irrelevant parts from globally estimated textures using saliency maps greatly helps to increase the discriminative capacity of BSIF-TOP features. The use of this feature could able to improve the performance of state-of-the-art shape and motion features by a good margin. In comparison to state-of-the-art and joint feature utilization methods discussed in Chapter 4, proposed method achieved superior recognition performance on low quality versions and subsets of three public datasets.

CHAPTER 6

CONCLUSION

This thesis presents a framework and two method for human activity recognition (HAR) in video, with low quality, i.e. low frame resolution, low frame rate, compression artifacts and motion blur. The problem of low video quality brings about more challenges to this domain. This thesis concerns two important areas of HAR in low quality videos as follows: (1) Spatio-temporal framework, (2) spatio-temporal feature representation.

Existing approaches available in activity recognition are mainly focused on good quality videos. They were mainly designed to process high quality videos by extracting rich feature sets and tediously process them with high computational cost. The strategy of the current methods are not suitable for the case of low quality videos due to poorness of visual information. The survey of existing literature in Chapter 2 gives an overview of recent approaches in HAR and gives a critique of how they are not appropriate for low quality videos.

The main goal of this thesis is to focus on exploiting various types of video features attainable in low quality video for better recognition of activities. One spatio-temporal feature based framework and two methods– joint feature utilization and mid-level feature bank, are proposed. The framework extracts three types of spatio-temporal features, namely shape, motion and texture, and uses textures. In comparison with existing frameworks that utilize only shape and motion features, the current framework adds textural features which is able to improve the recognition performance by a good margin but with a smaller computational cost.

The concluding remarks and future work for each of the proposed methods are described below:

The joint feature utilization method

The proposed joint feature utilization method uses shape-motion and textural features to improve the performance of action recognition in low quality videos. In comparison with current methods that mainly rely on only shape and motion features, the use of textural features is a novel proposition that improves the recognition performance by a good margin ($\approx$ 5-20% depending on datasets), as per the experimental results shown in Chapter 4. Though there is high relevancy of low quality videos in real-world application, but to date, there is no systemic work that investigates the problem of low video quality. This method also draws some interesting observations, (1) the trajectory based method is more sensitive to the spatial resolution than the interest point based method, (2) consideration of a large number of features for building codebook (using k-means algorithm) perform better only if dissimilar type of features are used i.e., HOG/HOF – HOG based on gradients and HOF based on optical flow features, (3) BoW encoding perform better if quality of video becomes very low, i.e., KTH-SD_4, while FV does opposite, and (4) the response of videos with the very poor quality i.e., KTH-SD_4, and HMDB51-BQ are found very strong across most of experimental settings with the inclusion of textures.

Spatio-temporal mid-level feature bank

The proposed spatio-temporal mid-level (STEM) feature bank integrates the advantages of local explicit patterns from either interest points and dense trajectories, and global salient statistical patches, in order to improve the activity recognition performance in low quality videos. The idea of pruning textures with 3D saliency patch, refereed as salient textures, removes the activity irreverent features from BSIF-TOP features. It helps to increase the discriminative properties of textures. In comparison to state-of-the-art, proposed method achieved superior recognition performance on various low quality versions and subsets created from three public datasets, as per the experimental results shown in Chapter 5.

6.1 Future Directions

The future direction for each of the proposed methods are described below:

The joint feature utilization method

There are various possible ways to extend this idea. The joint feature utilization idea relies on three spatio-temporal features- two based on local interest point features and local trajectories while the other based on global features. Since current interest point and trajectory methods do not consider the problem of low quality video, so design of methods with the concern of low quality problems might be a potential direction of work. The BSIF features features seem more appropriate with other textural features for low quality videos, so this is also a potential direction of further research. So far, only handcrafted features were used in this work, it would be interesting to see how unsupervised features perform with low quality videos.

Spatio-temporal mid-level feature bank

There are various possible ways of extending STEM. The shape and motion based features used in this method can be improved by pruning irrelevant features using saliency maps. The saliency approach used in STEM do not always provide accurate saliency map if videos with complex backgrounds are considered. So, a saliency map that is robust to the complex environments can be formulated to produce a more discriminative feature set. The saliency obtained by deep learning (R. Zhao et al., 2015; X. Li et al., 2015; Pan & Jiang, 2016; Kuen et al., 2016) can be a potential direction to solve this issue.

APPENDIX A

BAG OF VISUAL WORDS MODEL

Bag of Visual Words (BoVW), also known as Bag of Features (BoF) is a very popular method inherited from bag of words (BoW) model used in text and language processing that consider visual features as videos or images. The BoVW is very popular among computer vision researchers and the its efficiency and effectiveness have been demonstrated in massive visual classification tasks especially in large-scale human activity recognition tasks (H. Wang et al., 2009, 2011; H. Wang & Schmid, 2013; Peng et al., 2014).

Generally, the BoVW representation of a video comprises of three distinct steps: extraction of features, vector quantization of features, and feature histogram representation with pooling and normalization. The feature histograms are then classified by a classifier which is usually the non-linear support vector machine (SVM). The first step, 'feature extraction' deals with the extraction of features from video. The literature is packed with many feature extraction methods for activity recognition from video. Popular feature extraction method includes STIP (Laptev, 2005), Cuboid (Dollár et al., 2005), Hessian (Willems et al., 2008), Dense Sampling (H. Wang et al., 2009), Dense Trajectories (H. Wang et al., 2011), and Improved Dense Trajectories (H. Wang & Schmid, 2013). Feature extraction methods usually represent visual features in a form of vectors. Each video is then represented by its closest visual word from visual codebook constructed from feature vectors in 'vector quantization' step. Finally, to obtain a holistic representation of features 'pooling' is applied. The use of pooling usually create some sparsity in features, which is avoided by normalizing the features. The general pipeline for BoVW representation is shown in Figure A.1.

A.1 Generation of codebook

A codebook is a base element to describe a video, which is constructed from a set of feature vectors. In action recognition, there two popular methods for generating codebook: (1) hard partitioning where codebook is generated by partitioning the feature space into various informative regions called visual words or codewords, and (2) soft partitioning where a generative model capture the distribution of features in terms of probability. For the first method k-means is widely used and GMM (Gaussian mixture model) is widely used for the second method. While k-means only provides the mean of codewords, the GMM additionally provides the shape of the distribution. The description of both methods is given below:

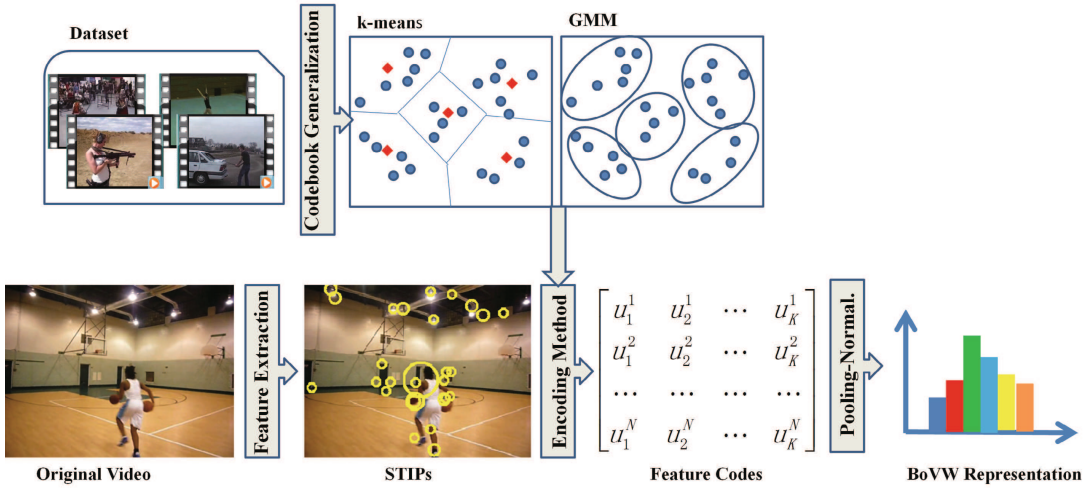


Figure A.1: The Bag of Visual Words (BoVW) method for human activity recognition. Figure reproduced from (X. Wang et al., 2013)

***k*-means**: There are various vector quantization method available in the literature such as spectral clustering, hierarchical clustering, and k-means clustering. The performance of k-means is better than other clustering methods, and many activity recognition method used it for visual codebook generations. Given, a set of features $\{f_1, f_2, ..., f_M\}, f_m \in \mathrm{R}^D$, our objective is to partition the set into K number of clusters $\{c_1, c_2, ..., c_K\}, c_k \in \mathrm{R}^D$. Let for every feature f_m, we assign a set of binary identifier $r_{mk} \in \{0,1\}$; if f_m is assigned to the cluster k then $r_{mk} = 1$ and $r_{mj} = 0$, and $k \neq j$. The objective function can be defined as:

$$\min F(r_{mk}, c_k) = \sum_{m=1}^{M} \sum_{k=1}^{K} r_{mk} \parallel f_m - c_k \parallel^2 \tag{A.1}$$

The objective is to minimize the function F [1] by finding values for $\{r_{mk}\}$ and $\{c_k\}$. Usually, the optimization of the function is done in the iterative procedure where every iteration deals with two successive steps that deals with successive optimization with respect to r_{nk} and c_k.

Gaussian mixture model (GMM): GMM is a generative method that describes the distribution in a feature space:

$$p(f;\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(f;\mu_k,\Sigma_k) \tag{A.2}$$

where K is the number of mixtures, $\theta = \{\pi_1, \mu_1, \Sigma_1, \pi_2, \mu_2, \Sigma_2, ..., \pi_K, \mu_K, \Sigma_K\}$ is the parameter for model, and $\mathcal{N}(f;\mu_k,\Sigma_k)$ is a Gaussian distribution with D-dimension. Given a feature set $F = f_1, f_2, ..., f_M$, the optimal GMM parameters are obtained by maximum likelihood $\ln p(F;\theta) = \Sigma_m \ln p(f_m;\theta)$. We utilized an iterative expectation maximization (EM) algorithm for this problem solving.

[1] The notation $\parallel . \parallel$ express the l_2-norm; i.e., $\parallel a \parallel = \sqrt{\sum_{i=1}^{D} a_i^2}$

A.2 Encoding Methods

In this section, two popular encoding methods, namely vector quantization (VQ) and fisher vector (FV) are described. Assume F be a set of feature descriptors of D-dimension extracted from a video, hence we can write F = $[f_1, f_2, ..., f_N] \in R^{D \times N}$. Given a codebook D with K number of words, i.e., D = $[d_1, d_2, ..., d_K] \in R^{D \times K}$, the goal of encoding is to calculate a code for f with D. The code vector is represented by the notation u_n, which is of dimension of D for vector quantization, and $2KD$ fisher vector (FV) representation, where K is the number of GMM clusters.

Vector Quantization (VQ): It is also popularly known as 'hard-assignment' encoding. Due to simplicity, VQ has been used by many researchers for recognizing human activities. In VQ, each feature descriptor f_n is represented by its nearest visual words in the codebook:

$$u_{nk} = \begin{cases} 1 & if \quad k = arg\ min_k \parallel f_n - d_k \parallel^2 \\ 0 & otherwise \end{cases} \tag{A.3}$$

Fisher Vector (FV): Fisher vector or kernel encoding recently got popularity for their excellent results in activity recognition, especially in large-scale recognition problems (H. Wang & Schmid, 2013). With the advantages of codebook based methods, FV also provide the advantage of using generative models. Let's say we have a generative model $p(f;\theta)$ in a feature space, and F = $\{f_1, f_2, ..., f_N\}$ is a set of N number of features extracted from a video. The video is then can be represented in a form of log likelihood or the gradient vector with respect to the parameters of the model (Jaakkola et al., 1999) expressed as:

$$G_\theta^{\mathrm{F}} = \frac{1}{N} \nabla_\theta \log p(F;\theta) \tag{A.4}$$

where, the dimension of vector is independent of number of features N, and only depends on the number of θ parameters. A natural kernel K on these gradients can be defined as:

$$K(F,I) = G_\theta^{\mathrm{FN}} F_\theta^{-1} G_\theta^{\mathrm{I}} \tag{A.5}$$

where F_θ is the fisher information matrix of $p(F;\theta)$ defined as:

$$F_\theta = \mathrm{E}_{F \sim p(F;\theta)}[\nabla_\theta \log p(F;\theta) \nabla_\theta \log p(F;\theta)^T] \tag{A.6}$$

The F_θ is a positive definite and symmetric so, we can defined the Fisher Vector as:

$$\mathscr{G}_\theta^F = F_\theta^{-1/2} G_\theta^F \tag{A.7}$$

Since we use GMM for estimation of $p(x;\theta)$ so, we assume that the covariance matrices Σ_k are diagonal. The FV encoding can be defined as:

$$\mathscr{G}^F_{\mu,k} = \frac{1}{N\sqrt{\pi_k}} \sum_{n=1}^{N} \gamma_n(k) \left(\frac{F_n - \mu_k}{\sigma_k} \right), \tag{A.8}$$

$$\mathscr{G}^F_{\mu,k} = \frac{1}{N\sqrt{\pi_k}} \sum_{n=1}^{N} \gamma_n(k) \left(\frac{(F_n - \mu_k)^2}{\sigma_k^2} - 1 \right) \tag{A.9}$$

where $\gamma_n(k)$ is the soft assignment of feature F_n to j-th Gaussian j:

$$\gamma_n(k) = \frac{\pi_k \mathscr{N}(F;\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi \mathscr{N}(F;\mu_k,\Sigma_k)} \tag{A.10}$$

The final gradient vector u is form by concatenating $\mathscr{G}_{\mu,k}$ and $\mathscr{G}_{\sigma,k}$.

A.3 Feature Pooling and Normalization

Given the feature coding coefficients, a pooling operation is used for obtaining holistic representation p from all local feature descriptors of a video. Two common strategies for pooling are found in literature (X. Wang et al., 2013):

Sum pooling: Using sum pooling , the i-th component of p can be defined as:

$$p_i = \sum_{n=1}^{N} u_{ni} \tag{A.11}$$

Max pooling: Using max pooling, the i-th component of p can be defined as:

$$p_i = \max(u_{1i}, u_{2i}, ..., u_{ni}) \tag{A.12}$$

Boureau, Ponce, and LeCun (2010) analyzed these two methods, and it shows that the max pooling is the more preferable sum pooling for sparse features. In order to obtain a more distributed arrangement of features many methods further normalize the pooled features. According to (X. Wang et al., 2013), there are three normalization strategies are available:

l_1-normalization: Using l_1 normalization method, the feature p is divided by its l_1-norm:

$$p = p / \sum_{i=1}^{K} \text{abs}(p_i) \tag{A.13}$$

l_2-normalization: Using l_2 normalization method, the feature p is divided by its l_2-norm:

$$p = p/\sqrt{\sum_{i=1}^{K} \text{abs}(p_i^2)} \tag{A.14}$$

Power normalization: Using power normalization method, we use the following function with both dimension:

$$f(p_i) = \text{sign}(p_i)\ \text{abs}(p_i)^{\alpha}, \quad 0 \leq \alpha \leq 1 \tag{A.15}$$

where α is normalization parameter. The power normalization is also combinable with l_1-normalization and l_2-normalization.

REFERENCES

[1] Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *Proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1597–1604).

[2] Aggarwal, J. K., & Cai, Q. (1997). Human motion analysis: A review. In *Proc. of Nonrigid and Articulated Motion Workshop* (pp. 90–102).

[3] Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, *43*(3), 16.

[4] Ahad, M. A., Tan, J., Kim, H., & Ishikawa, S. (2010). A simple approach for low-resolution activity recognition. *Int. J. Comput. Vis. Biomech*, *3*(1).

[5] Ahad, M. A. R., Ogata, T., Tan, J., Kim, H., & Ishikawa, S. (2008). A complex motion recognition technique employing directional motion templates. *International Journal of Innovative Computing, Information and Control*, *4*(8), 1943–1954.

[6] Ahad, M. A. R., Tan, J., Kim, H., & Ishikawa, S. (2011). SURF-based spatio-temporal history image method for action representation. In *Proc. of IEEE International Conference on Industrial Technology (ICIT)* (pp. 411–416).

[7] Ahsan, S. M. M., Tan, J. K., Kim, H., & Ishikawa, S. (2014). Histogram of dmhi and lbp images to represent human actions. In *Proc. of IEEE International Conference on Image Processing (ICIP)* (pp. 1440–1444).

[8] Ali, S., Basharat, A., & Shah, M. (2007). Chaotic invariants for human action recognition. In *Proc. of 11th International Conference on Computer Vision (ICCV)* (pp. 1–8).

[9] Aminian Modarres, A., & Soryani, M. (2013). Body posture graph: a new graph-based posture descriptor for human behaviour recognition. *IET Computer Vision*, *7*(6), 488–499.

[10] Babu, R. V., & Ramakrishnan, K. (2004). Recognition of human actions using motion history information extracted from the compressed video. *Image and Vision Computing*, *22*(8), 597–607.

[11] Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2011). Sequential deep learning for human action recognition. In *Human Behavior Understanding* (pp. 29–39). Springer.

[12] Baumann, F., Ehlers, A., Rosenhahn, B., & Liao, J. (2014). Computation strategies for volume local binary patterns applied to action recognition. In *Proc. of 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 68–73).

[13] Baumann, F., Ehlers, A., Rosenhahn, B., & Liao, J. (2016). Recognizing human actions using novel space-time volume binary patterns. *Neurocomputing*, *173*, 54–63.

[14] Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In *Proc. of 10th IEEE International Conference on Computer Vision (ICCV)* (Vol. 2, pp. 1395–1402).

[15] Bobick, A., & Davis, J. (1996). An appearance-based representation of action. In *Proc. of 13th International Conference on Pattern Recognition (ICPR)* (Vol. 1, pp. 307–312).

[16] Bobick, A. F. (1997). Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *352*(1358), 1257–1265.

[17] Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *23*(3), 257–267.

[18] Boureau, Y.-L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proc. of 27th International Conference on Machine Learning (ICML)* (pp. 111–118).

[19] Bradski, G. R., & Davis, J. W. (2002). Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, *13*(3), 174–184.

[20] Bregonzio, M., Gong, S., & Xiang, T. (2009). Recognising action as clouds of space-time interest points. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1948–1955).

[21] Bregonzio, M., Li, J., Gong, S., & Xiang, T. (2010). Discriminative topics modelling for action feature selection and recognition. In *Proc. of British Machine Vision Conference (BMVC)* (pp. 1–11).

[22] Cai, J., Merler, M., Pankanti, S., & Tian, Q. (2015). Heterogeneous semantic level features fusion for action recognition. In *Proc. of 5th ACM on International Conference on Multimedia Retrieval (ICMR)* (pp. 307–314).

[23] Cai, J. X., Feng, G.-c., & Tang, X. (2013). Human action recognition using oriented holistic feature. In *Proc. of 20th IEEE International Conference on Image Processing (ICIP)* (pp. 2420–2424).

[24] Cao, L., Luo, J., Liang, F., & Huang, T. S. (2009). Heterogeneous feature machines for visual recognition. In *Proc. of 12th International Conference on Computer Vision (ICCV)* (pp. 1095–1102).

[25] Chaaraoui, A. A., Climent-Pérez, P., & Flórez-Revuelta, F. (2013). Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, *34*(15), 1799–1807.

[26] Chakraborty, B., Holte, M. B., Moeslund, T. B., & Gonzàlez, J. (2012). Selective spatio-temporal interest points. *Computer Vision and Image Understanding (CVIU)*, *116*(3), 396–410.

[27] Chandrashekhar, V. H., & Venkatesh, K. (2006). Action energy images for reliable human action recognition. In *Proc. of Asian Symposium on Information Display (ASID)* (pp. 484–487).

[28] Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (n.d.). Return of the devil in the details: Delving deep into convolutional nets..

[29] Chen, C.-C., & Aggarwal, J. (2009). Recognizing human action from a far field of view. In *Proc. of Workshop on Motion and Video Computing (WMVC)* (pp. 1–7).

[30] Chen, C.-C., & Aggarwal, J. (2011). Modeling human activities as speech. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3425–3432).

[31] Chen, M., Gong, L., Wang, T., & Feng, Q. (2015). Action recognition using lie algebrized gaussians over dense local spatio-temporal features. *Multimedia Tools and Applications*, *74*(6), 2127–2142.

[32] Chen, X., Cheng, Y., & Yi, Y. (2015). Features extraction approach based on dense salient trajectories in videos. In *Proc. of International Symposium on Bioelectronics and Bioinformatics (ISBB)* (pp. 132–135).

[33] Cheng, G., Wan, Y., Saudagar, A. N., Namuduri, K., & Buckles, B. P. (2015). Advances in human action recognition: A survey. *arXiv preprint arXiv:1501.05964*.

[34] Cho, J., Lee, M., Chang, H. J., & Oh, S. (2014). Robust action recognition using local motion and group sparsity. *Pattern Recognition*, *47*(5), 1813–1825.

[35] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 1, pp. 886–893).

[36] Dawn, D. D., & Shaikh, S. H. (2015). A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector. *The Visual Computer*, 1–18.

[37] Dogan, E., Eren, G., Wolf, C., & Baskurt, A. (2015). Activity recognition with volume motion templates and histograms of 3d gradients. In *Proc. of IEEE International Conference on Image Processing (ICIP)* (pp. 4421–4425).

[38] Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance* (pp. 65–72).

[39] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2625–2634).

[40] Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Proc. of 9th IEEE International Conference on Computer Vision (ICCV)* (pp. 726–733).

[41] Evangelidis, G., Singh, G., & Horaud, R. (2014). Skeletal quads: Human action recognition using joint quadruples. In *Proc. of International Conference on Pattern Recognition (ICPR).*

[42] Everts, I., Van Gemert, J. C., & Gevers, T. (2014). Evaluation of color spatio-temporal interest points for human action recognition. *IEEE Transactions on Image Processing (TIP)*, *23*(4), 1569–1580.

[43] Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

[44] Folgado, E., Rincón, M., Carmona, E. J., & Bachiller, M. (2011). A block-based model for monitoring of human activity. *Neurocomputing*, *74*(8), 1283–1289.

[45] Fu, X., McCane, B., Albert, M., & Mills, S. (2013). Action recognition based on principal geodesic analysis. In *Proc. of 28th International Conference of Image and Vision Computing New Zealand (IVCNZ)* (pp. 259–264).

[46] Gaidon, A., Harchaoui, Z., & Schmid, C. (2012). Recognizing activities with cluster-trees of tracklets. In *Proc. of British Machine Vision Conference (BMVC)* (pp. 30–1).

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

[48] Gilbert, A., Illingworth, J., & Bowden, R. (2011). Action recognition using mined hierarchical compound features. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *33*(5), 883–897.

[49] Guo, K., Ishwar, P., & Konrad, J. (2010). Action change detection in video by covariance matching of silhouette tunnels. In *Proc. of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (pp. 1110–1113).

[50] Gupta, R., Jain, A., & Rana, S. (2013). A novel method to represent repetitive and overwriting activities in motion history images. In *Proc. of International Conference on Communications and Signal Processing (ICCSP)* (pp. 556–560).

[51] Han, J., & Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *28*(2), 316–322.

[52] Harandi, M. T., Sanderson, C., Shirazi, S., & Lovell, B. C. (2013). Kernel analysis on grassmann manifolds for action recognition. *Pattern Recognition Letters*, *34*(15), 1906–1915.

[53] Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Advances in neural information processing systems* (pp. 545–552).

[54] Harjanto, F., Wang, Z., Lu, S., Tsoi, A. C., & Feng, D. D. (2015). Investigating the impact of frame rate towards robust human action recognition. *Signal Processing*.

[55] Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Proc. of Alvey Vision Conference* (Vol. 15, p. 50).

[56] Herath, S., Harandi, M., & Porikli, F. (2016). Going deeper into action recognition: A survey. *arXiv preprint arXiv:1605.04988*.

[57] Hou, X., Harel, J., & Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *34*(1), 194–201.

[58] Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., & Huang, T. S. (2009). Action detection in com-

plex scenes with spatial and temporal ambiguities. In *Proc. of 12th International Conference on Computer Vision (ICCV)* (pp. 128–135).

[59] Ikizler, N., & Duygulu, P. (2009). Histogram of oriented rectangles: A new pose descriptor for human action recognition. *Image and Vision Computing*, *27*(10), 1515–1526.

[60] Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*(11), 1254–1259.

[61] Jaakkola, T. S., Haussler, D., et al. (1999). Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, 487–493.

[62] Jain, A., Gupta, A., Rodriguez, M., & Davis, L. (2013). Representing videos using mid-level discriminative patches. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2571–2578).

[63] Jain, M., Jégou, H., & Bouthemy, P. (2013). Better exploiting motion for better action recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2555–2562).

[64] Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *35*(1), 221–231.

[65] Jiang, Y.-G., Dai, Q., Xue, X., Liu, W., & Ngo, C.-W. (2012). Trajectory-based modeling of human actions with motion reference points. In *Proc. of European Conference Computer Vision (ECCV)* (pp. 425–438). Springer.

[66] Johansson, G. (1975). Visual motion perception. *Scientific American*.

[67] Kannala, J., & Rahtu, E. (2012). Bsif: Binarized statistical image features. In *Proc. of 21st International Conference on Pattern Recognition (ICPR)* (pp. 1363–1366).

[68] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1725–1732).

[69] Kataoka, H., Aoki, Y., Iwata, K., & Satoh, Y. (2015a). Evaluation of vision-based human activity recognition in dense trajectory framework. In *Advances in Visual Computing* (pp. 634–646).

Springer.

[70] Kataoka, H., Aoki, Y., Iwata, K., & Satoh, Y. (2015b). Evaluation of vision-based human activity recognition in dense trajectory framework. In *Proc. of 11th International Symposium on Visual Computing (ISVC)* (p. To Appear).

[71] Ke, S.-R., Thuc, H. L. U., Lee, Y.-J., Hwang, J.-N., Yoo, J.-H., & Choi, K.-H. (2013). A review on video-based human activity recognition. *Computers*, *2*(2), 88–131.

[72] Kellokumpu, V., Zhao, G., & Pietikäinen, M. (2008a). Human activity recognition using a dynamic texture based method. In *Proc. of British Machine Vision Conference (BMVC)* (Vol. 1, p. 2).

[73] Kellokumpu, V., Zhao, G., & Pietikäinen, M. (2008b). Texture based description of movements for activity analysis. In *Proc. of Int. Conf. on Computer Vision Theory and Applications (VISAPP)* (Vol. 1, pp. 206–213).

[74] Kellokumpu, V., Zhao, G., & Pietikäinen, M. (2011). Recognition of human actions using texture descriptors. *Machine Vision and Applications*, *22*(5), 767–780.

[75] Kim, T.-K., & Cipolla, R. (2009). Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *31*(8), 1415–1428.

[76] Kim, W., Lee, J., Kim, M., Oh, D., & Kim, C. (2010). Human action recognition using ordinal measure of accumulated motion. *EURASIP journal on Advances in Signal Processing*, *2010*(1), 1–11.

[77] Klaser, A., Marszałek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3D-gradients. In *Proc. of 19th British Machine Vision Conference (BMVC)* (pp. 275–1).

[78] Kovashka, A., & Grauman, K. (2010). Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2046–2053).

[79] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: a large video database for human motion recognition. In *Proc. of International Conference on Computer Vision (ICCV)* (pp. 2556–2563).

[80] Kuen, J., Wang, Z., & Wang, G. (2016). Recurrent attentional networks for saliency detection.

arXiv preprint arXiv:1604.03227.

[81] Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision (IJCV)*, *64*(2-3), 107–123.

[82] Laptev, I., & Lindeberg, T. (2003). Space-time interest points. In *Proc. of IEEE International Conference on Computer Vision (ICCV)* (pp. 432–439). IEEE.

[83] Laptev, I., Marszałek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).

[84] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

[85] Li, B., Ayazoglu, M., Mao, T., Camps, O. I., & Sznaier, M. (2011). Activity recognition using dynamic subspace angles. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3193–3200).

[86] Li, X., Zhao, L., Wei, L., Yang, M., Wu, F., Zhuang, Y., ... Wang, J. (2015). DeepSaliency: Multi-task deep neural network model for salient object detection. *arXiv preprint arXiv:1510.05484*.

[87] Lin, Z., Jiang, Z., & Davis, L. S. (2009). Recognizing actions by shape-motion prototype trees. In *Proc. of 12th International Conference on Computer Vision (ICCV)* (pp. 444–451).

[88] Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision (IJCV)*, *30*(2), 79–116.

[89] Liu, C., & Yuen, P. C. (2010). Human action recognition using boosted eigenactions. *Image and Vision Computing*, *28*(5), 825–835.

[90] Liu, J., Luo, J., & Shah, M. (2009). Recognizing realistic actions from videos "in the wild". In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1996–2003).

[91] Liu, L., Shao, L., Li, X., & Lu, K. (2016). Learning spatio-temporal representations for action recognition: A genetic programming approach. *IEEE Transactions on Cybernetics*, *46*(1), 158–170.

[92] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, *60*(2), 91–110.

[93] Lu, W.-L., & Little, J. J. (2006). Simultaneous tracking and action recognition using the pca-hog descriptor. In *Proc. of 3rd Canadian Conference on Computer and Robot Vision* (pp. 6–6).

[94] Ma, S., Bargal, S. A., Zhang, J., Sigal, L., & Sclaroff, S. (2015). Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *arXiv preprint arXiv:1512.07155*.

[95] Ma, S., Sigal, L., & Sclaroff, S. (2015). Space-time tree ensemble for action recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5024–5032).

[96] Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2929–2936).

[97] Mathieu, M., Couprie, C., & LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*.

[98] Matikainen, P., Hebert, M., & Sukthankar, R. (2009). Trajectons: Action recognition through the motion analysis of tracked features. In *Proc. 12th International Conference on Computer Vision Workshops (ICCV Workshop)* (pp. 514–521).

[99] Mattivi, R., & Shao, L. (2009). Human action recognition using lbp-top as sparse spatio-temporal feature descriptor. In *Proc. of Computer Analysis of Images and Patterns* (pp. 740–747).

[100] Meng, H., Pears, N., & Bailey, C. (2006). Recognizing heuman actions based on motion information and svm. In *Proc. of 2nd IET International Conference on Intelligent Environments (IE)* (Vol. 1, pp. 239–245).

[101] Messing, R., Pal, C., & Kautz, H. (2009). Activity recognition using the velocity histories of tracked keypoints. In *Proc. of 12th International Conference on Computer Vision (ICCV)* (pp. 104–111).

[102] Misra, I., Zitnick, C. L., & Hebert, M. (2016). Unsupervised learning using sequential verification for action recognition. *arXiv preprint arXiv:1603.08561*.

[103] Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, *104*(2), 90–126.

[104] Murakami, M., Tan, J. K., Kim, H., & Ishikawa, S. (2010). Human motion recognition using directional motion history images. In *Proc. of International Conference on Control Automation and Systems (ICCAS)* (pp. 1445–1449).

[105] Murthy, O., & Goecke, R. (2013). Ordered trajectories for large scale human action recognition. In *Proc. of IEEE International Conference on Computer Vision Workshops (ICCV Workshop)* (pp. 412–419).

[106] Murthy, O., & Goecke, R. (2015). Harnessing the deep net object models for enhancing human action recognition. *arXiv preprint arXiv:1512.06498*.

[107] Murthy, O. R., & Goecke, R. (2015). Ordered trajectories for human action recognition with large number of classes. *Image and Vision Computing*, *42*, 22–34.

[108] Nguyen, T. V., Song, Z., & Yan, S. (2015). STAP: Spatial-temporal attention-aware pooling for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *25*(1), 77–86.

[109] Ogata, T., Tan, J. K., & Ishikawa, S. (2006). High-speed human motion recognition based on a motion history image and an eigenspace. *IEICE TRANSACTIONS on Information and Systems*, *89*(1), 281–289.

[110] Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., ... Desai, M. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3153–3160).

[111] Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *24*(7), 971–987.

[112] Ojansivu, V., & Heikkilä, J. (2008). Blur insensitive texture classification using local phase quantization. In *Image and signal processing* (pp. 236–243). Springer.

[113] Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, *11*(285-296), 23–27.

[114] Päivärinta, J., Rahtu, E., & Heikkilä, J. (2011). Volume local phase quantization for blur-insensitive dynamic texture classification. In *Image Analysis* (pp. 360–369). Springer.

[115] Pan, H., & Jiang, H. (2016). A deep learning based fast image saliency detection algorithm. *arXiv preprint arXiv:1602.00577*.

[116] Peng, X., Qiao, Y., Peng, Q., & Qi, X. (2013). Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition. In *Proc. of British Machine Vision*

Conference (BMVC).

[117] Peng, X., Wang, L., Wang, X., & Qiao, Y. (2014). Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *arXiv preprint arXiv:1405.4506*.

[118] Perronnin, F., Sánchez, J., & Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *Proc. of European Conference on Computer Vision (ECCV)* (pp. 143–156). Springer.

[119] Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, *28*(6), 976–990.

[120] Qian, H., Mao, Y., Xiang, W., & Wang, Z. (2010). Recognition of human activities using SVM multi-class classifier. *Pattern Recognition Letters*, *31*(2), 100–111.

[121] Rahman, S., See, J., & Ho, C. (2015). Action recognition in low quality videos by jointly using shape, motion and texture features. In *Proc. of IEEE Int. Conf. on Signal and Image Processing Applications (ICSIPA)* (pp. 83–88).

[122] Rahman, S., See, J., & Ho, C. (2016). Deep CNN object features for improved action recognition in low quality videos. In *International Conference on Computational Science and Engineering (ICCSE)* (p. To appear).

[123] Ramana Murthy, O., Radwan, I., & Goecke, R. (2014). Dense body part trajectories for human action recognition. In *Proc. of IEEE International Conference on Image Processing (ICIP)* (pp. 1465–1469).

[124] Raptis, M., Kokkinos, I., & Soatto, S. (2012). Discovering discriminative action parts from mid-level video representations. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1242–1249).

[125] Raptis, M., & Soatto, S. (2010). Tracklet descriptors for action modeling and video analysis. In *Proc. of European Conference Computer Vision (ECCV)* (pp. 577–590). Springer.

[126] Reddy, K. K., Cuntoor, N., Perera, A., & Hoogs, A. (2012). Human action recognition in large-scale datasets using histogram of spatiotemporal gradients. In *Proc. of 9th International Conference on Advanced Video and Signal-Based Surveillance (AVSS)* (pp. 106–111).

[127] Reddy, K. K., & Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, *24*(5), 971–981.

[128] Rodriguez, M. D., Ahmed, J., & Shah, M. (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).

[129] Roh, M.-C., Shin, H.-K., & Lee, S.-W. (2010). View-independent human action recognition with volume motion template on single stereo camera. *Pattern Recognition Letters*, *31*(7), 639–647.

[130] Rubinstein, M., Liu, C., & Freeman, W. T. (2012). Towards longer long-range motion trajectories.

[131] Ryoo, M., Chen, C.-C., Aggarwal, J., & Roy-Chowdhury, A. (2010). An overview of contest on semantic description of human activities (SDHA) 2010. In *Recognizing Patterns in Signals, Speech, Images and Videos* (pp. 270–285). Springer.

[132] Sadanand, S., & Corso, J. J. (2012). Action bank: A high-level representation of activity in video. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1234–1241).

[133] Schüldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local SVM approach. In *Proc. of 17th International Conference on Pattern Recognition (ICPR)* (Vol. 3, pp. 32–36).

[134] Seo, H. J., & Milanfar, P. (2009). Nonparametric bottom-up saliency detection by self-resemblance. In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)* (pp. 45–52).

[135] Seo, H. J., & Milanfar, P. (2011). Action recognition from one example. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *33*(5), 867–882.

[136] Shao, L., Ji, L., Liu, Y., & Zhang, J. (2012). Human action segmentation and recognition via motion and shape analysis. *Pattern Recognition Letters*, *33*(4), 438–445.

[137] Shao, L., Zhen, X., Tao, D., & Li, X. (2014). Spatio-temporal laplacian pyramid coding for action recognition. *IEEE Transactions on Cybernetics*, *44*(6), 817–827.

[138] Shechtman, E., & Irani, M. (2007). Space-time behavior-based correlation-or-how to tell if two underlying motion fields are similar without computing them? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *29*(11), 2045–2056.

[139] Shi, F., Laganiere, R., & Petriu, E. (2015). Gradient Boundary Histograms for action recognition. In *Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1107–1114).

[140] Shi, F., Laganière, R., & Petriu, E. (2016). Local part model for action recognition. *Image and Vision Computing*.

[141] Shi, F., Petriu, E., & Laganiere, R. (2013). Sampling strategies for real-time action recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2595–2602).

[142] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems* (pp. 568–576).

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

[144] Srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). Unsupervised learning of video representations using lstms. *arXiv preprint arXiv:1502.04681*.

[145] Sultani, W., & Saleemi, I. (2014). Human action recognition across datasets by foreground-weighted histogram decomposition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 764–771).

[146] Sun, J., Mu, Y., Yan, S., & Cheong, L.-F. (2010). Activity recognition using dense long-duration trajectories. In *Proc. of IEEE International Conference on Multimedia and Expo (ICME)* (pp. 322–327).

[147] Sun, J., Wu, X., Yan, S., Cheong, L.-F., Chua, T.-S., & Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2004–2011).

[148] Sundaram, N., Brox, T., & Keutzer, K. (2010). Dense point trajectories by gpu-accelerated large displacement optical flow. In *Proc. of European Conference Computer Vision (ECCV)* (pp. 438–451). Springer.

[149] Tsai, D.-M., Chiu, W.-Y., & Lee, M.-H. (2015). Optical flow-motion history image (OF-MHI) for action recognition. *Signal, Image and Video Processing*, *9*(8), 1897–1906.

[150] Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(11), 1473–1488.

[151] Uijlings, J. R., Duta, I., Rostamzadeh, N., & Sebe, N. (2014). Realtime video classification using dense HOF/HOG. In *Proc. of International Conference on Multimedia Retrieval (ICMR)* (p. 145).

[152] Vedaldi, A., & Zisserman, A. (2012). Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *34*(3), 480–492.

[153] Vig, E., Dorr, M., & Cox, D. (2012). Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *Proc. of European Conference Computer Vision (ECCV)* (pp. 84–97). Springer.

[154] Vishwakarma, D., & Kapoor, R. (2015). Integrated approach for human action recognition using edge spatial distribution, direction pixel and-transform. *Advanced Robotics*, *29*(23), 1553–1562.

[155] Vishwakarma, S., & Agrawal, A. (2013). A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, *29*(10), 983–1009.

[156] Wang, C., Wang, Y., & Yuille, A. (2013). An approach to pose-based action recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 915–922).

[157] Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2011). Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3169–3176).

[158] Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision (CVPR)*, *103*(1), 60–79.

[159] Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In *Proc. of IEEE International Conference on Computer Vision* (pp. 3551–3558).

[160] Wang, H., Ullah, M. M., Klaser, A., Laptev, I., & Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *Proc. of British Machine Vision Conference (BMVC)* (pp. 124–1).

[161] Wang, H., & Yi, Y. (2015, Oct). Tracking salient keypoints for human action recognition. In *Proc. of IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (p. 3048-3053). doi: 10.1109/SMC.2015.530

[162] Wang, L., Hu, W., & Tan, T. (2003). Recent developments in human motion analysis. *Pattern recognition*, *36*(3), 585–601.

[163] Wang, L., Qiao, Y., & Tang, X. (2013). Motionlets: Mid-level 3d parts for human motion recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2674–2681).

[164] Wang, L., Qiao, Y., & Tang, X. (2014). Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge*, *1*, 2.

[165] Wang, L., Qiao, Y., & Tang, X. (2015). Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4305–4314).

[166] Wang, X., Farhadi, A., & Gupta, A. (2016). Actions ~ transformations. In *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (p. To appear).

[167] Wang, X., Wang, L., & Qiao, Y. (2013). A comparative study of encoding, pooling and normalization methods for action recognition. In *Proc. of Asian Conference on Computer Vision (ACCV)* (pp. 572–585). Springer.

[168] Wang, Y., & Mori, G. (2009). Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(10), 1762–1774.

[169] Wiegand, T., Sullivan, G. J., Bjøntegaard, G., & Luthra, A. (2003). Overview of the H. 264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, *13*(7), 560–576.

[170] Willems, G., Tuytelaars, T., & Van Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proc. of European Conference on Computer Vision (ECCV)* (pp. 650–663). Springer.

[171] Wu, Q., Wang, Z., Deng, F., Xia, Y., Kang, W., & Feng, D. D. (2013). Discriminative two-level feature selection for realistic human action recognition. *Journal of Visual Communication and Image Representation*, *24*(7), 1064–1074.

[172] Wu, X., Xu, D., Duan, L., & Luo, J. (2011). Action recognition using context and appearance distribution features. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 489–496).

[173] Wu, Z., Jiang, Y.-G., Wang, X., Ye, H., Xue, X., & Wang, J. (2015). Fusing multi-stream deep networks for video classification. *arXiv preprint arXiv:1509.06086*.

[174] Xu, H., Tian, Q., Wang, Z., & Wu, J. (2015). A survey on aggregating methods for action recognition with dense trajectories. *Multimedia Tools and Applications*, 1–17.

[175] Xu, X., Tang, J., Zhang, X., Liu, X., Zhang, H., & Qiu, Y. (2013). Exploring techniques for vision

based human activity recognition: Methods, systems, and evaluation. *Sensors*, *13*(2), 1635–1650.

[176] Yan, X., Chang, H., Shan, S., & Chen, X. (2014). Modeling video dynamics with deep dynencoder. In *Proc. of European Conference of Computer Vision (ECCV)* (pp. 215–230). Springer.

[177] Yang, Y., & Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1385–1392).

[178] Yeffet, L., & Wolf, L. (2009). Local trinary patterns for human action recognition. In *Proc. of 12th International Conference on Computer Vision (ICCV)* (pp. 492–497).

[179] Yi, Y., & Lin, Y. (2013). Human action recognition with salient trajectories. *Signal processing*, *93*(11), 2932–2941.

[180] Yuan, C., Li, X., Hu, W., Ling, H., & Maybank, S. (2013, June). 3D R transform on spatio-temporal interest points for action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[181] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4694–4702).

[182] Zhang, J. T., Tsoi, A. C., & Lo, S. L. (2014). Scale invariant feature transform flow trajectory approach with applications to human action recognition. In *Proc. of International Joint Conference on Neural Networks (IJCNN)* (pp. 1197–1204).

[183] Zhang, S., Yao, H., Sun, X., Wang, K., Zhang, J., Lu, X., & Zhang, Y. (2014). Action recognition based on overcomplete independent components analysis. *Information Sciences*, *281*, 635–647.

[184] Zhao, D., Shao, L., Zhen, X., & Liu, Y. (2013). Combining appearance and structural features for human action recognition. *Neurocomputing*, *113*, 88–96.

[185] Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *29*(6), 915–928.

[186] Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1265–1274).

[187] Zhou, F., & De la Torre, F. (2016). Generalized canonical time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 279–294.

[188] Ziaeefard, M., & Ebrahimnezhad, H. (2010). Hierarchical human action recognition by normalized-polar histogram. In *Proc. of 20th International Conference on Pattern Recognition (ICPR)* (pp. 3720–3723).

PUBLICATION LIST

Conference Proceedings

[1] Rahman, S., & See, J. (2016). Spatio-temporal mid-level feature bank for action recognition in low quality video. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1846–1850).

[2] Rahman, S., See, J., & Ho, C. (2016). Deep CNN object features for improved action recognition in low quality videos. In *International Conference on Computational Science and Engineering (ICCSE)* (p. To appear).

[3] Rahman, S., See, J., & Ho, C. C. (2015). Action recognition in low quality videos by jointly using shape, motion and texture features. In *Proc. of International Conference on Signal and Image Processing Applications (ICSIPA)* (pp. 83–88).

[4] Rahman, S., See, J., & Ho, C. C. (2017). Leveraging textural features for recognizing actions in low quality videos. In H. Ibrahim, S. Iqbal, S. S. Teoh, & M. T. Mustaffa (Eds.), *9th International Conference on Robotic, Vision, Signal Processing and Power Applications: Empowering Research and Innovation.* Singapore: Springer Singapore.

[5] See, J., & Rahman, S. (2015). On the effects of low video quality in human action recognition. In *Proc. of International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1–8).