HUMAN ACTIVITY CLASSIFICATION
USING
SPATIO-TEMPORAL FEATURE RELATIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

KUTALMIŞ AKPINAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

AUGUST, 2012

Approval of the thesis:

HUMAN ACTIVITY CLASSIFICATION USING SPATIO-TEMPORAL FEATURE RELATIONS

Submitted by **KUTALMIS AKPINAR** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen ____________________
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İsmet Erkmen ____________________
Head of Department, **Electrical and Electronics Eng.**

Assoc. Prof. Dr. İlkay Ulusoy ____________________
Supervisor, **Electrical and Electronics Eng. Dept., METU**

Examining Committee Members:

Prof. Dr. Uğur Halıcı ____________________
Electrical and Electronics Engineering Dept., METU

Assoc. Prof. Dr. İlkay Ulusoy ____________________
Electrical and Electronics Engineering Dept., METU

Assoc. Prof. Dr. Çağatay Candan ____________________
Electrical and Electronics Engineering Dept., METU

Assist. Prof. Dr. Fatih Kamışlı ____________________
Electrical and Electronics Engineering Dept., METU

Halil İbrahim Cüce, M.Sc. ____________________
Team Leader, ILTAREN TUBITAK

Date: ____________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Kutalmış AKPINAR

Signature :

ABSTRACT

HUMAN ACTIVITY CLASSIFICATION USING SPATIO-TEMPORAL FEATURE RELATIONS

Akpınar, Kutalmış
M.Sc., Department of Electrical and Electronics Engineering
Supervisor: Assoc. Prof. Dr. İlkay Ulusoy

August 2012, 62 pages

This thesis compares the state of the art methods and proposes solutions for human activity classification from video data. Human activity classification is finding the meaning of human activities, which are captured by the video. Classification of human activity is needed in order to improve surveillance video analysis and summarization, video data mining and robot intelligence. This thesis focuses on the classification of low level human activities which are used as an important information source to determine high level activities.

In this study, the feature relation histogram based activity description proposed by Ryoo et al. (2009) is implemented and extended. The feature histogram is widely used in feature based approaches; however, the feature relation histogram has the ability to represent the locational information of the features. Our extension defines a new set of relations between the features, which makes the method more effective for action description. Classifications are performed and results are compared using feature histogram, Ryoo's feature relation histogram and our feature relation histogram using the same datasets and the feature type. Our experiments show that feature relation histogram performs slightly better than the feature histogram, our feature relation histogram is even better than both of the two. Although the difference is not clearly observable in the datasets containing periodic actions, a 12% improvement is observed for the non-periodic action datasets. Our work shows that the spatio-temporal relation represented by our new set of relations is a better way to represent the activity for classification.

Keywords: Human Activity, Feature Based, Feature Relations, Histogram Based, Cuboid Feature.

ÖZ

ZAMANSAL VE YÜZEYSEL ÖZELLİK İLİŞKİLERİNİ KULLANARAK İNSAN HAREKETLERİNİN SINIFLANDIRILMASI

Akpınar, Kutalmış
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Anabilim Dalı
Tez Yöneticisi: Doç. Dr. İlkay Ulusoy

Ağustos 2012, 62 sayfa

Bu tez, videoda insan aktivitelerinin sınıflandırılması problemine önerilen çözümleri karşılaştırır ve yeni çözümler önerir. İnsan aktivite sınıflandırması; videoya çekilmiş olan insan aktivitelerinin anlamlarının bulunmasıdır. İnsan aktivitelerini sınıflandırmak; gözetleme videolarının analiz ve özetlenmesi, video veri madenciliği ve robot zekası konularında ilerleme kaydetmek için gerekmektedir. Bu tez yüksek seviyedeki insan aktivitelerinin tanınmasında önemli bir bilgi kaynağı olarak kullanılan düşük seviyeli insan hareketlerinin sınıflandırılması konusuna odaklanmıştır.

Bu çalışmada, Ryoo (2009) tarafından önerilmiş olan özellik ilişkisi histogramı bazlı aktivite tanımlayıcısı uygulandı ve genişletildi. Özellik bazlı yaklaşımlarda yaygın olarak kullanılan özellik histogramı ile karşılaştırırsak; özellik ilişkisi histogramı özelliklerin pozisyonal bilgisini de temsil etmektedir. Bizim önerdiğimiz geliştirme, yöntemi daha yetenekli kılan yeni bir özellik ilişkisi kümesidir. Özellik histogramı, Ryoo'nun özellik ilişkisi histogramı ve bizim özellik ilişkisi histogramımız, aynı özellik türü ve aynı video kümeleri ile sınıflandırmada kullanılmış ve sonuçlar karşılaştırılmıştır. Deneylerimiz özellik ilişkisi histogramının özellik histogramından biraz daha iyi olduğunu, bizim özellik ilişkisi histogramımızın ise ikisinden de daha iyi olduğunu göstermektedir. Fark periyodik olayları içeren video kümelerinde net bir şekilde gözlemlenemese de, periyodik olmayan olayları içeren video kümelerinde %12 lik bir gelişme gözlemlenmektedir. Çalışmamız, önerilen yeni ilişki kümesini kullanan özellik ilişkisi histogramlarının aktivite tanımlamada daha iyi bir yöntem olduğunu göstermektedir.

Anahtar Kelimeler: İnsan Aktivitesi, Özellik Bazlı, Özellik İlişkisi, Histogram Bazlı, Prizmasal Özellik.

To My Parents.

ACKNOWLEDGEMENTS

The author would like to express his sincere gratitude to his supervisor Assoc. Prof. Dr. İlkay Ulusoy for her guidance, insight, encouragement and criticism through the research.

The author would like to thank Dr Piotr Dollàr for sharing his work on cuboid features.

The assistance of Ms. Chris Taylor for proofreading is gratefully acknowledged.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3D, 2D, 1D	3 Dimensional, 2 Dimensional, 1 Dimensional
MEI	Motion Energy Image
MHI	Motion History Image
DTW	Dynamic Time Warping algorithm
LTI	Linear Time Invariant
SVM	Support Vector Machine
HMM	Hidden Markov Model
DBN	Dynamic Bayesian Network
CHMM	Coupled Hidden Markov Model
PCA	Principal Component Analysis
knn	K-nearest Neighborhood Classifier
SIFT	Scale-Invariant Feature Transform
HOF/HOG	Histogram of Optical Flow/ Histogram of Gradients
HOG3D	3-Dimensional Histogram of Gradients
Harris3D	3-Dimensional Harris feature descriptor
pLSA	Probabilistic latent semantic analysis

pLSA-ISM	Probabilistic Latent Semantic Analysis Implicit Shape Model

CHAPTER 1

INTRODUCTION

The increasing use of surveillance camera systems has created an urgent need for the automation of case detection from video data. Security camera systems are employed for surveillance purposes such as saving event records, intrusion detection, theft prevention, traffic control, detection of need for assistance from people who are drowning or have fallen, suspicious person and forgotten or suspicious luggage detection. They are employed in many kinds of public areas and special security regions. Given the amount of cameras in use, only a small number can be watched by a human being. Furthermore, employing people to watch surveillance cameras is expensive and would not be considered if the application does not have a crucial importance. Even if people are employed to watch cameras for detecting specific events, due to the psychological and biological limitations of human attention, there is a limit to the effectiveness of this process. The automation of event detections from surveillance cameras facilitates the efficient use of these cameras.

There are many studies that have investigated surveillance camera applications. For example; there are intrusion detection products designed to satisfy the needs of industry [1] and unaccompanied luggage detection has been studied so that methods can assist development of products [2]. These advances are due to the maturity level of computer vision methods including; background subtraction, motion and feature tracking. Although there are solutions to some of the case detection problems, most cases require improvements in computer vision.

The recognition of human activities from video data is an important area of computer vision research which aims to solve many automation problems in surveillance systems. Human activity recognition finds the meaning of human activities captured by video through a process of classification. Activities can be low level such as running, walking, hand clapping, bending, falling [3] or higher level activities like stealing, suspicious wandering, fighting, seeking help, meeting or making a presentation [4]. The actors can be group of people or individuals.

The aim of research into human activity recognition is not only concerned with surveillance camera systems, it is also useful for the development of robot intelligence, human-computer interface, video content analysis and data mining.

Computer vision methods on foreground subtraction, motion tracking, feature tracking are some important information sources used for human activity recognition, but they might not contain enough information to interpret meaning or intention of activity. In order to provide sufficient data for the recognition of human activities, new information sources such as lower level activity occurrences or motion features can be useful.

Human activities are classified in terms of their complexities. Most surveillance scenarios of human activity like stealing or interaction with others are considered to be complex level activities in human activity recognition research [4]. Methods developed for this level of activities are more likely to fulfill the industrial needs, however, the performance of those methods are limited by its information sources. Lower level human activities like running, bending or hand movements are major part of the source of information thus, the recognition and classification of lower level human activities is also an important area of human activity recognition research.

In this work, the main focus will be on lower level human activities in terms of complexity. The aim is to develop a method for classification of basic human activities so that it provides information for higher level human activity recognition methods and the experimental setups and datasets are also selected for this purpose. Rather than extracting human activities from video recordings, the experiments use human activity which is already extracted. Our experiments classify the extracts which include single or periodic occurrence of a human activity.

The classification of lower level human activities can be undertaken using various methods including; space-time volumes, space-time trajectories, space-time features and probabilistic or sequential models. Those methods differ in terms of their dependence on lower level processes, their ability to cope with changes in activity model or the activity types which they are able to model.

Space-time approaches model human activity in terms of the volumetric information in XY-T. The *space-time volume* approach represents human activity in its XY-T template. The t*rajectory based* approach uses body part trajectories to represent human activity and the *feature based* approach uses volumetric features in order to represent this activity. There are some major drawbacks to these methods such as the *space-time volume* approach not having much tolerance in the changes in activity model and the *trajectory based* approach usually requires body part extraction as a post process.

On the other hand, action can also be represented in terms of its states and sequential or probabilistic models are employed in order to model human activity. This type of modeling is also able to support more complex activities however; the disadvantage is their inability to cope with changes in the activity model or the requirement for a wide range of training set.

This study is based on feature based human activity recognition methods which select feature points from video extract which represents the event. These points are used for the event modeling. The classification of human activities using space-temporal features is a promising approach which has approach has almost no dependence on post processing procedures and it has a high ability to cope with changes from activity model.

Although, among the feature based methods there are some popular feature extraction processes used, how the feature information is to be combined for activity representation is still an issue. Composing the features with their spatio-temporal locations for action description is being studied in recent years [4]. The sequential representations of features describe events in strict timing requirements and feature histograms are not able to include the XY-T location information of the features. For this study, the spatio-temporal relation histogram method proposed by Ryoo et al. [5] is considered to be an important method in terms of activity representation.

In the research, several human activity representations for classification including spatio-temporal feature histograms and Ryoo's [5] spatio-temporal feature relation histograms were compared. An extension is proposed to Ryoo's method so that feature relations are set in XY-T rather than XY and T separately. The feature relation set was replaced with a more descriptive set. In the tests, the popular cuboid feature type was used in the event description as proposed in [6].

This thesis is organized as follows: Chapter 2 presents existing human activity classification methods in categories and analyzed feature based methods in a general structure; Chapter 3 outlines and compares the implemented methods, Chapter 4 gives a definition of the experimental setups and the results, Chapter 5 contains the concluding remarks on the implemented methods.

CHAPTER 2

LITERATURE SURVEY

Human activity recognition research proposes advancements in several fields. Some examples of those fields are surveillance video analysis [2], video data mining [7], video content search [8], human-computer interface [9], [10], and robot intelligence. Some examples of applications are analysis for airport, hospital or street surveillance cameras [2] for unaccompanied luggage, theft or accidents, patient monitoring; giving commands to computers; training artificial intelligence from video databases [11].

In this chapter, the previous researches will be examined according to their methodology but the application specific details will not be covered. Most of the studies covered in this chapter are related to methods which propose a general solution to the human activity classification concerning the activity complexity level. In Section 2.1, the related studies will be presented and categorized, then in Section 2.2 there a detailed analysis will be given on feature based methods and their common structures.

2.1. Overview of Human Activity Recognition Methods

Human activity recognition aims to detect and classify activities at various levels. These can be divided into 4 levels; *gestures*, *actions*, *interactions* and *group activities*. *Gestures* are atomic level meaningful components of human motion like *stretching arm, raising leg*. *Actions* are meaningful activities enacted by a person which can be composed of several or periodic gestures such as *walking*, *waving* and *punching*.

Interactions are human activities that involve two or more people or the interaction of a person with an object such as *two people fighting* or *stealing suitcase*. Finally, *group activities* are those performed by a conceptual group of people for example; *a group of people having meeting* or *a group of people marching* are some examples of group activities. [4]

There are several schemes used for the classification of activity recognition methods. Functionality based taxonomy was used by [12] and [13]. They categorized the methods according their structural layout according to basic levels of human activity analysis. The basic levels of activity recognition in their work were initialization[1], tracking, pose estimation and recognition. In [14], they classified methods as top-down and bottom-up.[2] Top down methods usually use geometric body reconstruction which separates the trajectory of each body element and bottom up methods utilize the low level features of action volume. In [4], Aggarwal et al. undertook an approach based on taxonomy. They categorized methods according to how the method describes the activity. Since their categorization covers the widest range of activity recognition methods, in the current work we present event recognition methods according to Aggarwal's categorization. According to their taxonomy, current activity recognition methods can be divided into three; space-time volume approaches which define the activity as 3D XYT volume, sequential approaches which define the activity as a sequence of observations and finally hierarchical approaches which define the activity as a combination of activities in smaller levels. A simplified hierarchy chart is given in Figure 1 and a more detailed version can be seen in [4].

2.1.1. Space-Time Approaches

One approach for human activity recognition is to consider human action as a 3D space time volume. Human action as a space-time volume is the set of pixels of the human located in video frame X-Y axis and the time axis. The typical methodology in this

[1] Initialization: Mainly concerns camera calibration, adoption of scene characteristics and model initialization [13].

[2] Top down: dividing into separate components to have more insight. Bottom up: synthesis of information.

approach is to construct a 3D space-time representation for each activity from training videos. 3D space-time representation of the new activity is also extracted and compared with previous representations. The similarities are calculated and the highest similarity is claimed to be the related activity class.

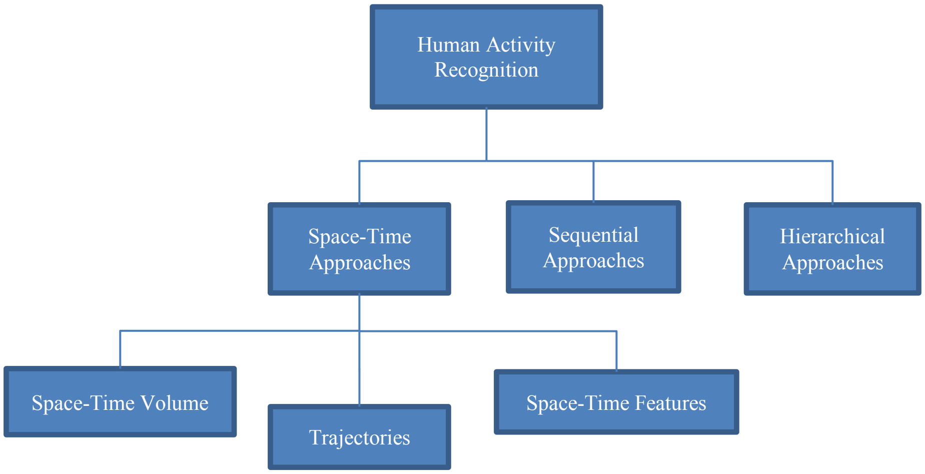


Figure 1: Approach based taxonomy of this review

There are several types of space-time methods with the use of shape information of space-time volume being the most basic one. Other space-time approaches include use of trajectories, and space-time features. The trajectories are tracking results of feature or body parts. Space-time features are similar to the features in object recognition methods, but they are in 3D.

Using shapes and the contents of 3D space-time volume is a direct approach to represent an event in a video. 3D space-time volumes are represented with models and comparison is undertaken using a distance or similarity calculation. Methods vary in terms of the model and comparison type. Template matching is one of the popular methods used in this approach [4].

In [15], Bobick represented 3D space-time volume with a pair of 2 dimensional images. One of those images is the Motion Energy Image (MEI) a binary image that holds a cumulative history of areas which are part of the motion areas. The Motion History

Image (MHI) is the image that holds the depth information about how recently the motion occurred in the related pixel of the frame. In MHI, the recently moving frames are brighter whereas older frames fade. For each action type, a statistical model of the MEI and MHI are created. A scale and translation invariant template matching method is used.

In [16], Shechtman estimated flows from a correlation of local patches around the video and used correlation values as a template for the activity. In his method, local patches of activity in a training video are correlated with the action volume and a correlation map is extracted. The action description is defined in this correlation map around the action. For the detection of similar actions the similarities between the correlation maps are evaluated.

Ke [17] used over-segmented videos in his work and the space-time volume is evaluated as smaller volumetric features. Classification searches are undertaken to find actor volumes among the activity. [4]

A disadvantage of the space-time volume representation is its vulnerability to changes from the action model. Space-time volume models may not show similarities between models of same action classes when the body pose, volume size or timing varies.

Another way of representing action in XY-T axis is to use trajectories of motion; these are the curves of specific body points in XY-T or XYZ-T space. Trajectories are useful in effectively representing the motions of body parts. They are extracted by continuous body part detection or feature tracking.

Sheikh used trajectories as set of space-time body part locations [18]. The view and actor variation among the same action class is eliminated by finding an appropriate linear space that holds common features between occurrences. This allows the detecting of action directly from the 2D detection of body parts.

Trajectories are valuable in their efficient representation of body motion. However, there is an issue with the extraction of the trajectories and research into body-part extraction and tracking is ongoing. Action recognition through trajectories requires too much post processing and is vulnerable to errors from these processes.

A space-time volume action can also be represented from its feature points. The interest points of an event are selected in an unsupervised manner and processed in order to represent an action. This method is covered in detail in 2.2 below.

2.1.2. Sequential Approaches

Other than space-time volume, a more dynamic representation of human action is to describe action as a sequence of observations. In this approach, the action is divided into several human poses. The sequences of human poses are learned and compared with the new human pose sequence for the classification. By expressing an action as sequence of poses, more complex human actions can be represented and recognized.

A direct approach to recognize sequences is to use a human pose sequence as a representation of the action. New video data is converted into sequences and categorized into one of the training sequences.

Darrell [19] and Gavrila [20] used Dynamic Time Warping (DTW) algorithm to compare training and test sequences [4]. This algorithm measures similarity between two sequences which may vary in time or speed. It is used in speech recognition and other signal processing applications [19]. Gavrila [20] extended the DTW approach applying it to an action model in XYZ. Lublinerman [21] modeled sequences of human poses as Linear Time Invariant (LTI) systems in which sequences of trajectories are converted to LTI system parameters and classified using a Support Vector Machine (SVM) [4].

Another approach for sequential representation of human action is the state models. From training set, models like the Hidden Markov Model (HMM) or Bayesian networks are trained in order to generate common model for the same kind of action sequences. This method requires a wide set of training sequences in order to support changes in activity model.

For the probabilistic state models, several types of HMMs are implemented. Yamato [22] and Starner [10] used a standard HMM which trains a model per activity. Starner [10] used this method to recognize American Sign Language. They used background estimation and tracking for post processing. In order to model human interactions which are not supported by previous HMM based methods, Oliver [23] used a coupled HMM (CHMM) and Park [24] used DBN. In Natarajan [25], they also extended CHMM in order to model waiting states of human beings. [4]

In most cases, sequential approaches use body poses as sequence elements and this requires a post processing. Although body pose is an easy feature to extract, it may not always be possible in the situation where there is occlusion with different objects. When occlusion occurs, the body pose cannot be estimated, and sequence is cut or the estimation may lead to a different result. In addition, while the direct comparison of sequences is not reliable in the case of variations in model, the use of probabilistic state models requires the training of too many samples in order to overcome variations in the model. [4].

2.1.3. HIERARCHICAL APPROACHES

For the recognition of more complex events such as stealing, suspicious wandering, group activities, human-object interactions, the action descriptors mentioned above will not suffice. This kind of event can be modeled in a hierarchical manner in which action is represented in terms of sub-events. Sub-events are obtained from the single layered approach mentioned above or from some other information obtained from processes such as tracking, background subtraction or object recognition.

Since this thesis does not focus on the detection of higher level activities, the hierarchical approaches are not evaluated however; our method of implementation presented here forms a good basis for hierarchical activity recognition which requires detection of atomic activities [5].

2.2. Feature Based Human Activity Recognition

One way of representing 3D action volume is to use features which are specific set of points and areas of 3D action volume. The type of points are extracted are expected to include decisive atomic action areas such as the joints in the action of knee bending or arm stretching, tracking facial parts in head movement or the fingers on arm stretching [6]. By clustering those points, we can use their occurrence relation as a description of the action type.

Due to their success in object recognition and image registration, feature based methods are considered to be promising for use in activity recognition [4]. For the feature extraction process of activity recognition, some feature types that are used in object recognition and their 3D types have already been studied [26] [27]. There are also feature types which depend specifically on temporal information in order to capture motion features.

Feature based methods are usually composed of 3 processes; feature extraction, action representation with features, and comparison with the new event. The layout for the feature based classification framework is given in Figure 2.

In the feature extraction process, usually a 3D kernel filter is applied to the selection of feature points. The video information around feature point is processed in order to obtain a useful feature description. Gradients, optical flows and the Principal Component Analysis (PCA) around the 3D video block are some popular processing methods used to obtain the feature description which are combined in order to represent related action. Creating statistics from the feature types or their relative positions are the methods that relate feature descriptors with actions. Then those statistics are compared in order to test whether a new event belongs to a particular class of event. The k-nearest neighborhood and SVMs are popular methods for categorization.

2.2.1. FEATURE EXTRACTORS AND FEATURE DESCRIPTORS

Feature extraction and feature description are used to select feature points from the video data and create vector representations for each feature. In the *feature extraction* process, decisive parts of the activity are specified. In the *feature description* process, the video information around the feature is converted into a common useful vector representation.

The aim of the *feature extraction* is to select areas of the activity video where motions of descriptive components occur. Atomic motions like knee bending, arm stretching and head turning are some examples of motions which are detected in the feature extraction process. However, the feature extraction process is not motion type specific. In most feature extraction processes, instead of the detection of each body part and its type of motion, a calculation method is generated so that all candidates of motions are labeled as features. This process is also called *interest point selection*.

Apart from *interest point selection*, *feature extraction* also includes extraction of video block around interest points. In this work, terms of *feature extraction* and *interest point selection* will be used interchangeably.

After extraction of features around interest points, the feature descriptors are created which are vector representations for feature points. The preparation of the feature representation process is to extract the useful information in the feature blocks and eliminate other information so that clustering and pattern recognition methods can be effectively applied to those vectors. Taking gradients [6] [27] [11] [28], gradient direction normalization [27], intensity and variance normalization [6], creation of intensity [6] or gradient histograms [11] [28] are some of the methods used for this purpose. Feature descriptors are usually compressed using PCA [6] and clustered using k-means [6] [27] [11] [28] [5] or hierarchical clustering. For action recognition, Laptev [26] extended the commonly used feature detectors of Harris [29] to 3D and he used a separable kernel to extract 3D corners.

In [6], Dollàr used windowed pixel values around the feature point which he called “cuboid”. The feature points are selected with a separable response function. After smoothing the video frames, 1D Gabor filters are applied in temporal axis. These Gabor filters are used in order to amplify areas which have motion fields with patterns. In Gabor filtering, periodic oscillations of intensity over time give the highest response. The complex motion of a body part which has a pattern may generate this kind of oscillation. The interest points are selected from the local maximas of response function output. The volume around this interest point is a cuboid.

In order to be used for event description, cuboids should be classified. Dollàr [6] calculated a feature descriptor from each cuboid. Vector of normalized intensity values or gradient values are used as the feature descriptor which are clustered using k-means.

As a feature descriptor, Scovanner [27] used a 3D SIFT descriptor which is an extension of the 2D SIFT descriptors proposed in [30] for object recognition and image registration. For the detection of feature points, as in [6], he used local maximas in response to a Gabor function. For the 3D SIFT descriptor, he used directional gradients around a feature volume. Gradient magnitudes are entered into a histogram according to gradient directions. The feature descriptor is obtained by rotating the directional gradient histogram so that the dominant directions come to same axis. Rotation ensures the directional invariance of feature descriptor and the descriptors are again clustered using k-means.

2.2.2. Action Descriptors

In the training process, after the extraction of spatio-temporal features from the video data, the features are used in order to define the action. The action descriptor describes the event occurrence. Depending on the type of action description, event matching is used for the classification of a current video to a specific class of event.

In feature based methods, sometimes the spatial and temporal location of the feature is omitted from action representation. This is called the *bag of words* approach. Regardless of their spatial and sometimes even the temporal position, the features of an event clip are clustered and counted in order to use them as an activity representation. Although this approach defines some view invariance and tolerance against temporal delays, it also results in a large amount of information loss which may cause confusion between activities with similar type of features.

Dollàr [6] and Scovanner [27] both use the bag of words approach. After clustering the features, they are placed into histograms to represent the action. A set of histograms are created from the set of video data in training process. These histograms are sometimes normalized to the frame or feature count.

Another approach for feature based action recognition is to use features with their spatio-temporal positions. In recent years, this approach has attracted an increasing amount of interest [4]. Unlike the approach of bag of features, these methods attempt to model the spatio-temporal locational information of the extracted features.

Wong et al., [31] extended basic *pLSA* used for activity recognition in [32], with an implicit shape model (*pLSA-ISM*). Their model captures the spatio-temporal location information of features relative to the center of the activity. [4]

Savarese et al., [33] proposed a method for capturing spatio-temporal proximity information among features. For each action video clip, they measure feature co-occurrence patterns in a local 3-D region and construct histograms called *ST correlograms*. [4]

In [5], they represented human action with the space-temporal relationship of features. The features of an event are extracted and clustered using k-means. Relations between features are placed in a relation histogram which is a 3D histogram with *featuretype x featuretype x relationship* number of bins. It counts the number of occurrences of a relation between two specific types of features. The relations used are *equals, before, meets, overlaps, during, starts*, and *finishes* in time and *near, x-near, y-near, far* in space. The event is recognized from the intersection of histograms. Rather than periodic events, this method claims recognition on short and non-periodic atomic events such as arm stretch and withdraw.

2.2.3. Feature Detection Methods For Feature Based Human Activity Recognition

Laptev's 3D Harris Interest Point Detector [26]:

In order to obtain a 3D feature type, Laptev [26] expanded the Harris feature detector [29] to 3D this detects locations where the image values have significant variations in spatial axes. Such interest points of a spatial image I^{sp} can be found from the second moment matrix integrated in the scale $\mu_i^2 = s\mu_l^2$;

$$\mu^{sp} = g^{sp}(\,\cdot\,;\ \mu_i^2\,) * \begin{pmatrix} \left(L_x^{sp}\right)^2 & L_x^{sp}L_y^{sp} \\ L_x^{sp}L_y^{sp} & \left(L_y^{sp}\right)^2 \end{pmatrix}$$

where L_x^{sp} and L_y^{sp} are the Gaussian derivatives of image I^{sp} in spatial directions;

$$L_x^{sp}(\,\cdot\,;\sigma_l^2)\ = \partial_x(g^{sp}(\,\cdot\ ;\sigma_l^2) * I^{sp})$$

$$L_y^{sp}(\,\cdot\,;\sigma_l^2)\ = \partial_y(g^{sp}(\,\cdot\ ;\sigma_l^2) * I^{sp})$$

and g^{sp} is the spatial Gaussian kernel.

Harris [29] shows that eigenvalues λ_1 and λ_2 of μ^{sp} represent the characteristic variations of image I^{sp} in both directions. To detect such points, the selection of the positive local maximas of the corner function below is proposed;

$$H^{sp} = det\left(\mu_{sp}\right) - k\ trace^2(\mu^{sp}) = \lambda_1\lambda_2\ -\ k\ (\lambda_1 + \lambda_2)^2$$

For the detection of space-time 3D interest points, Laptev defines an interest point as; the points where the video pixel value variations are significant in all 3 directions. From this definition, Laptev solves same problem for the 3D case. The second moment function for 3D case is given as;

$$\mu = g(\,\cdot\,;\ \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_y & L_yL_t & L_t^2 \end{pmatrix}$$

and the Harris corner function is defined as;

$$H = det(\mu) - k\ trace^3(\mu) = \lambda_1\lambda_2\lambda_3 - k\ (\lambda_1 + \lambda_2 + \lambda_3)^3$$

where λ_1 , λ_2 , λ_3 are eigenvalues of μ.

To detect space-temporal interest points, positive local maximas of the corner function H are taken [26].

In his experiments, Laptev used a scale selection method. The neighborhoods of the interest points are selected and clustered for use in the action description. In his work, he shows that the features are a match between training and testing videos of a person walking. Laptev's feature extractor is used in [3] with SVMs and they show that method is able to recognize actions from the Kth dataset explained in 4.1.2.

Dollàr's Interest Point Detector [6]:

The cuboid detector proposed by Dollàr [6] uses Gabor filtering in temporal domain in order to detect space temporal interest points. Dollàr suggests that temporal features have a higher importance than spatial ones and Gabor filtering is an appropriate way to detect them. Moreover, Gabor filtering in temporal domain will not give rise to regions not containing spatial features or one directional translational motion. The local maximas of the response function is taken as interest points.

The response function that Dollàr used in order to select interest points has the form;

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

where $g(x, y\ ; \mu)$ is the 2D Gaussian smoothing kernel and h_{ev} , h_{od} are quadrature pair of 1D Gabor filters applied temporally. The Gabor filters are defined as;

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)\ e^{-t^2/\tau^2}$$

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)\ e^{-t^2/\tau^2}$$

where $\omega = 4/\tau$. The filters are separable in all 3 axes. Two parameters σ and τ roughly specify the scale of the detector.

Since Dollàr's interest point selector is used in our implementation, method is explained in detail in section 3.1.

Dense Sampling for Interest Point Selection [34]:

In his comparative work, Wang [34] proposes an interest point selection method which involves the dense sampling extracts features at regular positions and scales in space and time. Unlike many of the feature extraction methods which extracts features in a specific scale, this method also includes a dense sampling in scale. As a result, there are 5 dimensions to sample from; (x, y, t, σ, τ) where x and y are spatial dimensions, t is the temporal dimension, σ and τ are the spatial and temporal scales respectively.

When spatial and temporal sampling is carried out with 50% overlap, scaling samples are taken with a $\sqrt{2}$ multiplication of scale parameters σ and τ. In his work, Wang compares several interest point selection and descriptor models, and shows that his work is especially successful for UCF sports datasets which have variety in terms of event scales [34].

2.2.4. Feature Descriptors For The Feature Based Human Activity Recognition

Dollàr's Cuboid Feature Descriptor [6]:

Dollàr [6] used the Cuboid descriptor he proposed with his interest point detector. His descriptor uses a video block around an interest point 6 times larger than the feature scaleσ. He compared several representations of video block such as histogram, gradient, optical flow. In conclusion, he preferred gradient values within the video block as his descriptor. In the training period, Dollàr calculated a PCA transformation in order to shrink the size of the gradient value vector. After the PCA, the vectors of gradient values are classified using k-means.

Of the feature description methods Dollàr's [6] descriptor is a widely used [35]. In the comparative study [34], they show that its performance is similar to the recently proposed methods in [11], [28] which are more complex in both computation and implementation. In the comparison, the achievement of the Cuboid descriptor is remarkable because it does not even support a scalar variance among features as in other methods. Since Dollàr's feature descriptor is used in the implementation, the method is explained in detail in 3.2.

3D SIFT Descriptor [27]:

In order to obtain a 3D feature type, in [27] 2D SIFT descriptor defined in [30] was expanded to a 3D case. In the area of image registration and object recognition SIFT descriptors are widely used. For the 2D SIFT descriptor in [30] they extract common feature points from image pyramids in order to select interest points which are immune to scale changes. Their directions are also omitted by selecting a common dominant direction from gradient histogram of the feature points. The histogram of gradient values which are brought to a common dominant direction is used as feature descriptors.

In [27], a new 3D type of SIFT descriptor is proposed. This method does not propose an interest point selection process as in the 2D SIFT of [30], but as a feature descriptor, he uses a histogram of gradients which points to a common dominant gradient direction. As for the interest point selection, they use the local maximas of Gabor response function as in [6].

In order to construct a feature descriptor, they calculate gradient values in the neighborhood of each interest point. The gradient values are collected in a histogram which counts gradient directions. The author used parallels and meridians to divide gradient directions into bins. From the gradient direction histogram, the dominant direction is specified. In order to construct the descriptor, each gradient histogram is transformed in space so that dominant gradient directions come to a common direction. For the normalization of dominant gradient direction, a 3D transformation for the histogram bin is proposed. The transformed histograms are clustered using k-means as in the other descriptors [6] and [26]. In study of 3D SIFT, several scales for interest point detection and histogram direction resolution are tested. Actions are described using histogram of clustered descriptors and SVMs.

The 3D SIFT method [27] is an imitation of SIFT descriptor in terms of feature description. On the other hand, since it does not involve a scale invariant interest point selection method, it cannot support scale invariance. Features can be considered view invariant since they involve the normalization of gradient direction.

The idea of the histogram of gradient orientations as a descriptor which is extension of SIFT descriptor is also implemented in other work. [11] Introduced the Histogram of gradients/Histogram of Optical Flow (HOF/HOG) descriptors and in [28], they introduced HOG3D around the same period of time (2008). Although their feature extraction methods and the handling of scale issues are different, in terms of the information type they are using, methods are similar. In the work presented in [6], histogram of gradient values was also included for comparison with several methods, but their implementation was not detailed in terms of the gradient direction normalization and scale handling.

2.2.5. ACTION DESCRIPTORS FOR THE FEATURE BASED HUMAN ACTIVITY RECOGNITION

Histogram of Features as Action Descriptor:

One of the most popular methods and the most basic approaches for describing human activity using space-temporal features is to use a histogram of the feature descriptors. The histogram of feature descriptors is a *bag of words* approach which does not use the space-temporal positions of the features, but it holds statistical information of how many features have occurred in an activity. It holds these action descriptor clusters as histogram bins.

A histogram of features is also easy to compare and create criteria for classification. K Nearest Neighborhood (knn) and Support Vector Machines (SVM) are some of the popular methods used for classification of descriptors [3].

The histogram of features is a widely used method for object recognition. For human activity classification, it is used especially for the studies which introduce new feature types [6] [27] [11] [28]. However, its success rate is limited in activity recognition. Nowadays, methods which include space-temporal position information on activity description are gaining popularity [4].

Space-Temporal Feature Relation Histogram for Activity Recognition [5]:

In [5], a method is proposed for the detection, classification and localization of human activities using spatial and temporal position relations between features. Their method holds a histogram of feature relations and uses the histogram intersection for the detection and classification of activities.

As in other feature based methods, the features of events are extracted and clustered using k-means. Relations between the features are placed in a 3D relation histogram which has *featuretype x featuretype x relationship* number of bins. It counts the number of occurrences of a relation between a specific feature pair. The relations used are *equals, before, meets, overlaps, during, starts*, and *finishes* in time and *near, x-near, y-near, far* in space. The event is recognized from the intersection of histograms. Rather than periodic events, this method claims recognition on short and non-periodic atomic events like arm stretch and withdraw.

This method is implemented in the current work and explained in section 3.4.

2.2.6. Classifiers For The Feature Based Human Activity Recognition

Support Vector Machine (SVM) for Activity Classification [3]:

A popular method for the classification of feature based activity models which are usually histograms, is to use SVM. This is a state-of-art margin classifier which has gained popularity within pattern recognition applications [3]. It is a non-probabilistic, linear classifier which is trained in supervised manner.

SVM is a supervised learning method for the classification of vectors which resides in some space. Given two set of training vectors with their classes, the SVM method calculates a linear transformation of these vectors to a space. In that space, a hyperplane exists such that it separates the two sets of training vectors with a gap that is as wide as possible. The transformation and the hyperplane are calculated in the training process for the two sets of classes. For classification of test data, the test feature vector is mapped to that space. The class is determined according to the side that the hyperplane mapped vector resides in [36]. A calculation procedure for linear transformation and hyperplane is given in [3].

CHAPTER 3

THE SPATIO-TEMPORAL FEATURE RELATIONS METHODOLOGY

In this study, a feature based human activity classification method based on the framework in 2.2.1 is implemented. The focus for the selection of each process was to compare the feature histogram method, the feature relation histogram method proposed by Ryoo [5] and the proposed feature relation histogram with the new set of relations. For the feature type, Dollàr's cuboid features are used and for the classifier, the knn classification is used. As an alternative to the knn classifier, Ryoo's histogram comparison method proposed for event detection was also implemented and its classification performance is evaluated.

Although a general framework is valid for the feature based methods, there are various possible combinations of feature types, activity descriptors and classification methods. Some feature types such as Dollàr's cuboid [6], Laptev's 3D Harris [26] and the HOG/HOF descriptors [28] have already gained popularity in feature based activity recognition studies. Classifiers are a well-studied topic of Pattern Recognition. On the other hand, of all these processes the activity descriptor is a bottleneck. The only well-known method for activity descriptor is the feature histogram method which is not a promising one due to its insufficient information content to describe action.

The current work explores and improves the newly proposed method, feature relation histogram for activity description. The use of spatio-temporal feature relations is promising since it includes the flexibility of histogram based methods whilst retaining the space-temporal location information. This method is claimed to be appropriate for the description of atomic non-periodic activities. As a contribution to the feature relation method, new types of feature relation sets and histogram types are proposed which have a better ability to represent actions.

As a basis for Ryoo’s method, a 3D feature type is used. Ryoo's method offers the flexibility to use any 3D feature type. For the feature extractor and feature descriptor, cuboid features were used as proposed in [6]. According to the information from the literature survey, the cuboid feature is the most commonly used feature type in recent activity recognition research. It enables the comparison between the proposed action descriptor in this study with others. The comparison of interest point selectors and feature types in [34] also shows that the cuboid feature type’s performance is similar to the other state-of art methods such as HOF/HOG or HOG3D. Some of those state-of-art interest point selectors like dense sampling, Hessian [34], are also not appropriate for the current study since they sample features in a 5 dimensional space which includes scale factors. In the current study, the cuboid feature method is used as they are implemented within cuboids toolbox (4.2) by Dollàr.

The method is applied parallel to the layout in Figure 2. Some parameters are trained from a sample subset of the training set. Using trained parameters along with the method, action descriptors are extracted from the test and training sets. Descriptor vectors are classified using the proposed classifier.

The method starts with the selection of useful interest points from the activity video clips. The interest point selection process utilizes the selection of motion fields which also have spatial features. After selecting those types of points, the video blocks around those points are used for processing. Around each interest point, a video block is extracted which is called as the *feature cuboid*.

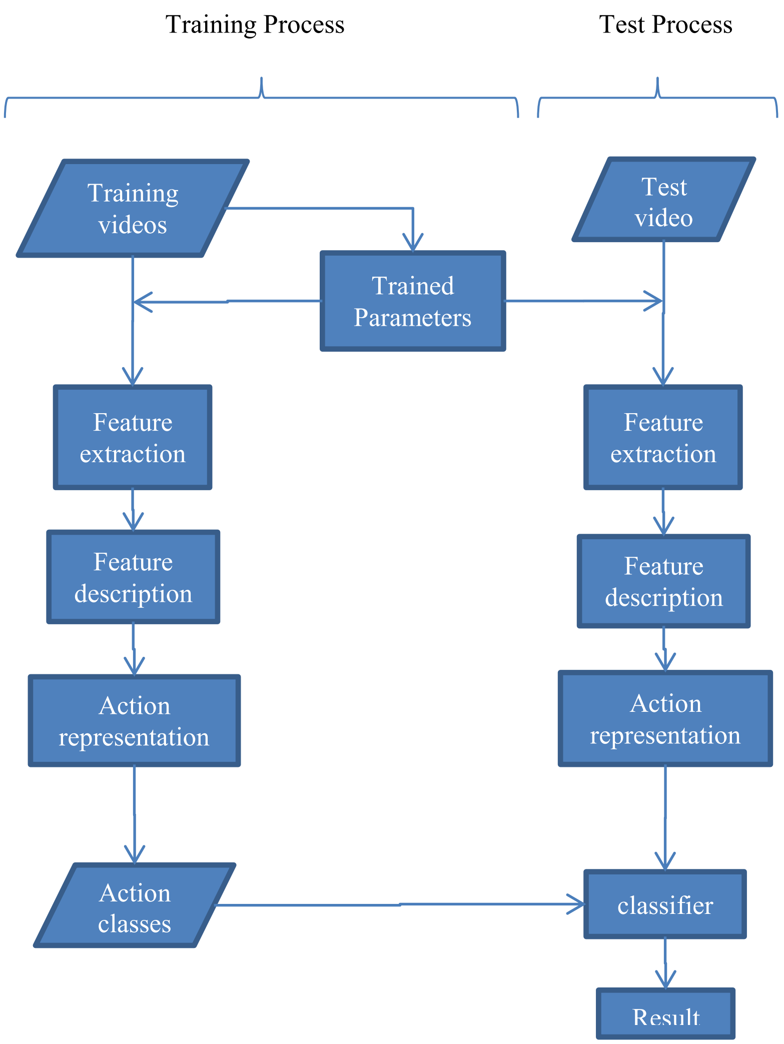


Figure 2: Classification Process

Feature cuboids as video blocks are not sufficiently useful to be used in action description; they need not be optimized in terms of useful information. Rather than intensity images, gradients are calculated and used since gradients show motion features better than intensity video blocks. To eliminate the noise level dimensions before clustering those gradient blocks are compressed to the dominant bases using PCA. From a sample training set the features are evaluated and a useful linear transformation is calculated using PCA. The insignificant dimensions are removed then the compressed features are clustered and indexed using k-means clustering. After feature extraction, a list of features is obtained with its class, spatial and temporal occurrence intervals.

Using cuboid feature descriptors extracted from the action clip, feature relation histogram based action descriptors are constructed. A list of relations between each feature pair is extracted for the video clip. These relations are composed into a histogram to represent action. For classification of event descriptors, a knn classifier with chi-squared distance criteria is used. In addition to chi-squared distance, a detection algorithm proposed in [5] was tested for classification purposes.

In this chapter, the methods used in this study are explained in detail. The method is composed of the cuboid interest point selector, the cuboid feature descriptor, action descriptors and classification of action descriptors.

3.1. Interest Point Selection

The aim of the interest point selection is to select the areas of the video clips where motions of descriptive components occur. Motions such as knee bending, arm stretching and head turning are examples of motions that are to be captured in the interest point selection process. In order to be able to cover unlimited kinds of activities and descriptive parts, the interest point selection process should not be motion type or component type specific. Instead of the detection of each component and its type of motion, all spatial features undergo complex translational motion are considered as interest points.

For the ease of calculation, interest points are defined as; spatial features undergoing complex motion which is defined as the motion type that creates a signature of the shape of the moving object and the motion type. That signature must be observed from small video block around the event. Corners or edges which reverse direction, moving parts with a specific temporal or spatial frequency are most favorable candidates for interest points.

A popular approach for the selection of interest points is to apply a descriptive kernel filter and select the local maxima of the response function as interest points. In terms of kernel filters, the Laplacian of Gaussian, Gabor [14] [6] and 3D Harris [26] feature detectors are some of the methods used for 3D motion interest point detection. Our interest point detector is based on Dollàr's method in [6]. Our descriptive kernel is composed of different separable linear filters on spatial and temporal axes.

On the temporal axis Gabor filtering is applied. This gives high responses to temporal oscillations which are created by motion of periodic spatial features or the periodic motion of spatial features. Periodic motions, turn of directions, passing several feature edges after each other which might be result of complex motions will give high responses. On the other hand, edges undergoing pure translational motion are not expected to give high responses as in the LOG filter case. Areas without a spatial feature will not give a response either even if they undergo a motion.

On the spatial axis, Gaussian smoothing is applied in order to increase the predictability of the system and to select a spatial frequency in which features will be detected. The response function energy of the Gabor filter is highly dependent on the sharpness of the edges. Normalization of edges is required in order to obtain results that are independent of this sharpness. The normalization level also selects the scale of the feature extraction method. If smoothing is high, then feature selection is sensitive to low frequency features otherwise the feature selection is sensitive to high frequency features. This normalization is obtained by applying Gaussian smoothing in spatial axis. This method also prevents irregularities on sharpness of the edges due to aliasing on the captured image. Those intensity variations on the edges create irrelevant local maximas on the Gabor filter response and they need to be removed.

With horizontal and spatial filters, the response function has the form:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

$g(x, y\,; \mu)$ is a 2D Gaussian smoothing kernel applied in spatial directions. h_{ev} and h_{od} are quadrature pair of 1D Gabor filters applied in temporal direction. These are defined as:

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)\, e^{-t^2/\tau^2}$$

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)\, e^{-t^2/\tau^2}$$

As in Dollàr's implementation, $\omega = 4/\tau$ is used. Parameters σ and τ are given roughly as the spatial and temporal scale of detector [6]. Filters are applied in separable manner. First, $I * g$ is calculated, then the odd and even Gabor filters are applied to the smooth image. R is calculated by combining odd and even filter outputs. Convolution with g can also be applied as separable filters in x and y axes.

As interest points, we seek for the local maximas of response function. The response function is usually expected to give its local maximum when a variation in the direction of the object occurs, but the translational motions of some objects with periodic features can also create ridge-like local maximal areas. Especially for those ridge-like places, local maximas are evaluated in a neighborhood. A radius is used in order to select only one maxima point around some neighborhood. Only the maximum of the local maximas in that neighborhood are considered as an interest point. The radius size is selected proportional to the cuboid size. A sample for the filter response and point selection is given in Figure 3.

3.2. Feature Descriptor

After the calculation of interest point locations cuboids [6] are extracted around each interest point. These video blocks are used as sources of information about these points. Cuboid is simply the pixel values around the interest point in a spatial and temporal window. Selected cuboids should contain most of the feature which cause the local maxima in response function, so the size is calculated proportional to the scale of the interest point detector filters. For the response function equations given above, spatial scale is σ and temporal scale is τ. Cuboid sizes are chosen as 6 times those scale values [6]. Cuboids are centered at interest points.

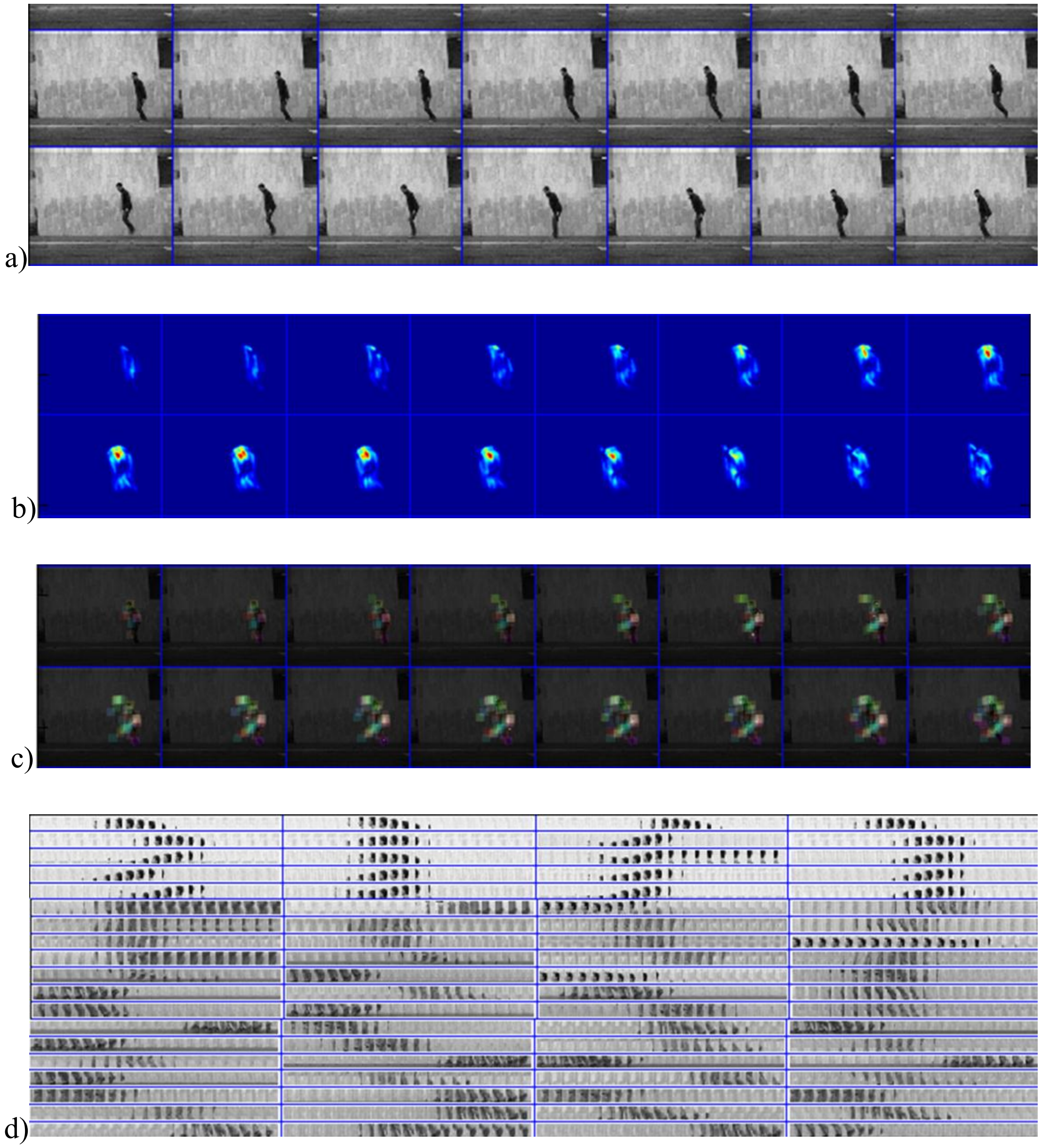


Figure 3: Feature extraction from video.
a) A cut from jump action of the Weizmann Dataset
b) Response function to the interest point selector kernel
c) Selected interest points and cuboids in rectangle
d) cuboid features extracted from video

A useful vector representation of a cuboid and a distance measure between those vectors are required. Some useful vector representations are, pixel values, gradient values, histogram of gradients, flow fields and 3D SIFT which was proposed by [27]. Feature expressions of cuboids which are ready to be compared for classification are called feature descriptors [34].

Dollàr [6] studied several Euclidean distance measurements for comparison. His selection of descriptor type is motivated from previous 2D and 3D feature types. In the current study, gradient based cuboid comparison is used since Dollàr obtained the highest success ratio in that type of vector. In the gradient based method, the gradient of current intensity field on the cuboid is used. A 3D gradient vector is obtained for each pixel in cuboid. As feature descriptor, the flattened values of gradient vectors from 3 dimensions are used. The distance is calculated as the Euclidean distance between feature descriptors.

To extract a cuboid, a cuboid shaped intensity field around interest point is taken. The gradient vector for each pixel in a cuboid is defined as;

$$\begin{bmatrix} G_x(i,j,k) \\ G_y(i,j,k) \\ G_t(i,j,k) \end{bmatrix} = \begin{bmatrix} I(i,j,k) - I(i+1,j,k) \\ I(i,j,k) - I(i,j+1,k) \\ I(i,j,k) - I(i,j,k+1) \end{bmatrix}$$

where G_x, G_y, G_z are three channels of the gradient and I is the intensity cuboid.

A Gaussian smoothing kernel is applied to each of the gradient channels in order to normalize magnitudes. Several Gaussian kernels are selected in order to cover different scales. The resulting equation is;

$$G_{x,y,t} = \begin{bmatrix} g_1 * G_x \\ g_1 * G_y \\ g_1 * G_t \\ g_2 * G_x \\ g_2 * G_y \\ g_2 * G_t \\ \vdots \end{bmatrix}$$

where $G_{x,y,t}$ is the Gaussian smooth result and $g_1(x, y, z\,; \sigma_1, \tau_1)$, $g_2(x, y, z\,; \sigma_2, \tau_2)$ are the Gaussian smooth kernels with different scale values. Using 1 or 2 scale values would be sufficient.

In order to reduce the dimensionality of $G_{x,y,z}$, PCA is used which reduces the dimension of the descriptor by transforming it to a lower dimensional linear space with minimum information loss.

3.3. Feature Classes

The interest point detector in the current study is expected to give a response to specific types of features which will be used in the proposed event descriptor. On the other hand, these feature descriptors can be of an infinite type because of the span of its space. It is necessary to map the feature descriptors to classes in order to make appropriate use of them.

The interest point detector proposed in the current work prevents some kind of features from rising therefore, the number of classes for our feature descriptor types can be practically useful. Although, the appearance of the same activities may vary greatly because of the subject or manner differences, in the proposed method, they are expected to give rise to similar features. Due to the coherence between similar activities and the motion ability of human body parts, it is expected that the useful feature descriptors can be assigned to some classes.

In order to create classes of feature descriptors, a large random set of feature descriptors is gathered from relevant scenarios within the same scale. Since a large set of feature descriptors strengthens the validity of feature classes, training for feature classes is undertaken in the training process. After training, a map is constructed from feature descriptors to feature classes. This map is used for mapping a new feature descriptor to a feature class.

In the proposed method, k-means clustering is used to create feature classes. K-means is applied to feature descriptors whose dimensions are reduced using PCA. Euclidean distance is used as the distance measure between descriptors. K-means clustering defines clusters as centroids. *k* number of centroids are created such that distance from the centroids to the elements are at their minimum and each element is assigned to the nearest centroid.

3.4. Event Descriptor

The set of pairwise relations between features gives a compact representation of the 3D feature occurrence information in a video block. This set of relations is not significantly affected by differences between occurrences.

Pairwise relations between features are collected in a histogram of relations in order to construct action description. For each ordered feature pair, the applicable relations are extracted. The number of relations can be 1 or 2 depending on the preferred relation set for the method.

The proposed action description is the set of histograms extracted from training videos. In order to construct an action description from the training videos, we extract histogram of feature relations around each video block in which the related action occurred. A set of those histograms are used as the action description consisting of *N* histograms of size *featuretype x featuretype x number of relations* where *N* is the number of training videos concerned with the action type.

3.4.1. RYOO'S SPATIO-TEMPORAL RELATION HISTOGRAM

In Ryoo's original study, one temporal relation and one spatial relation is generated for each ordered feature pair (a, b). In the current study the relationship histogram holds the number of occurrences of a type of relation between each type of feature pair. Thus, the number of bins in our histogram is featuretype x featuretype x number of relationships.

Two histograms; one for the temporal relations and the other for the spatial relations are held separately.

For the temporal relations, Ryoo uses the set of temporal predicates and spatial predicates shown below;

Allen's Temporal Predicates

Allen's temporal relations between feature A and feature B are as shown below;

Equals: the starting and ending times are equivalent.

Before: *feature A* ends before *feature B* starts

Meets: *feature B* starts when *feature A* ends

Overlaps: *feature B* starts while *feature A* is in process. Systematically, the start of *feature B* is between the start and end of *feature A*.

During: *feature A* starts and ends while *feature B* is in process.

Starts: They start simultaneously, *feature A* ends first.

Ends: *feature A* starts first, *feature A* and *feature B* end simultaneously.

In addition to these relations there are also the inverse of those relations except the *equals* relation.

In terms of point wise features, only the *equals*, *before* are appropriate to be considered. In terms of features which are fixed in size, only equals, before, overlaps and their inverse relations if applicable are considered. In the current case, since cuboid features are fixed in size, the relations are for the features fixed in size.

Spatial Predicates

Relations between the features in the spatial XY axes are near, x-near, y-near and far. Those definitions rely on some distance threshold in terms of pixel values. Distances are measured between the centers of features. Definitions of predicates are;

Near: Euclidean distances between XY coordinates of features are below the threshold.

X-near: They are not *near*, but the distances between x coordinates are below the threshold.

Y-near: They are not *near* or *x-near*, but the distances between y coordinates are below the threshold.

Far: They are not *near*, *x-near* or *y-near*.

3.4.2. 3D FEATURE RELATION HISTOGRAM

Ryoo's method holds spatial and temporal relations in different histograms or histogram bins. This method ignores relevance between the spatial and temporal relations. On the other hand, an action is more meaningful if described with words "*Feature A coming after feature B, around same coordinates*" or "*Feature A coming after feature B around same horizontal coordinate, but a different vertical coordinate*".

In this study, a new set of relations is proposed in order to describe actions. Each of the new relation type describes a temporal relation together with a spatial relation.

During the test on Ryoo's relations (4.3), it was also discovered that including far relations in histograms is not a good idea if they do not represent a meaningful pair. In the feature relation sets used in this study, an upper limit is applied to the feature relation distances. For the temporal upper bound, maximum motion cycle of the target events was considered. In terms of the upper limit of the spatial distance, maximum spatial size of an event cycle was considered.

The new relation set is defined as the Cartesian product of the temporal and spatial feature sets which are similar to those in Ryoo's method except that they do not include far relations above a certain distance. In the current case, the relation types for fixed size of features are used.

In the current study the feature relation histogram has *featuretype x featuretype x number of spatial relations x number of temporal relations* number of bins. Since there are 3 temporal and 4 spatial relations in the implementation, the number of bins increases from 7N to 12N. The histogram bins are filled according to the following rule; feature pairs are considered for the spatial and temporal relations; for each feature pair, 1 feature relation is generated if they are above the relevance thresholds for distance; else, no relation is generated.

3.5. Activity Detection and Classification

After the preparation of the action descriptors for each video in the training and test set, the action descriptor of the test video is classified into an action class using the action descriptors of training videos. For this process, the classifiers used for Pattern Recognition can be implemented. In the current method, the knn classifier is used.

Given a test vector to be classified and training vector sets for a set of classes, the distances from test vector to all training vectors are calculated. By considering those distances as neighborhood criteria, the k numbers of most near training vectors are selected. The most common class among those k vectors is selected for the class assignment.

In the case of the current study, the test vector is the action descriptor of the test video clip and the training classes are action descriptors of the training videos. Descriptors are the histogram of the features explained above. As for the distance measure, two different methods are used; the Chi-squared distance and the spatio-temporal relation match proposed by Ryoo for detection of events in [5]. This second method is explained in

3.5.1. In the method in this study k=1 was used which considers only the nearest neighbor.

A good choice for distance measure in the application in the current study is the Chi-squared distance. This is a useful distant measure for a histogram comparison in the Computer Vision applications. Given n dimensional histograms with v and w as vectors, this Chi-squared distance measure is given as;

$$d(v,w) = \frac{\sum_{i=1}^{n} \frac{(v_i - w_i)^2}{v_i + w_i}}{2}$$

3.5.1. Spatio-Temporal Relation Match

For the detection of events from random video sequences, Ryoo [5] proposed a method for a similarity measure. Detection is considered if similarity is detected over a certain threshold and the optimal threshold is obtained from the training set. In the method proposed in this thesis, this similarity measure is used as a distance measure in the knn classifier in order to test its classification performance. Obtaining the classification performance of a detection method is a good way to measure possibility of false detection.

The similarity measure between the histograms depends on the intersection of the histograms. Between the training video and testing video histograms, the common occurrences of feature relations are counted. This is done by taking minimums between histogram bin pairs and summing them up. If a video clip contains sufficient common relations of actions with the training video, it is considered to include an event from a related action type. This comparison method is able to detect existence of an event type even from a video clip that includes several types of events.

In order to find a common threshold value for the similarity measure to be considered as detection, a training set can be used. In order to use a common threshold for the similarity, the similarities should be normalized to an interval 0-1 by dividing similarity by the total number of features in training histogram.

Given histograms v from the test video clips, and the w for the training video, the similarity measure is given as;

$$s(v, w) = \frac{\sum_{i=1}^{n} \min(v_i, w_i)}{\sum_{i=1}^{n} w_i}$$

In order to use the similarity measure for classification purposes $d = 1/s$ is taken as the distance measure between histograms.

CHAPTER 4

THE EXPERIMENTS

The experimental setup, aims to validate and evaluate human activity classification ability of feature relation histogram based methods explained in Chapter 3. The experiments include analysis of the effective parameters, comparison of several algorithms under the same setups, and the performance results of suggested improvements.

The experiments began with extraction of features as explained in sections 3.1, 3.2 and 3.3. Cuboid features were used in feature histograms as in the original work [6]. Ryoo's relation histogram method was also implemented using the same feature sets. This work also has been tested on standard datasets and the results are available in the literature. Ryoo's method is analyzed in terms of weak points and improvements are suggested, implemented, tested and compared. In the current study, the performance measure was the classification ability for short video clips which included a single occurrence of some type of event. The reference was the performance of the feature histogram method where the original implementation was used.

The standard datasets present in the literature were preferred in order to make comparisons with existing studies. For the first results during the implementation of the algorithms, Weizmann dataset was used due to its appropriate quality and small video count; this allows clear debug of the results. Then, KTH dataset was used due to its popularity in literature. For the third results, the dataset proposed in the original work of spatio-temporal feature relations [5] was used.

In this chapter, an explanation of the datasets, experimental setups and important parameters are given. The results will be evaluated and compared.

4.1. Datasets

In the experiments, popular and appropriate datasets from the literature were used for the classification setups. All the datasets used consist of short video clips which include periodic or a single occurrence of an activity type by a single actor. The same activity was performed by different actors to create a variety of clips for the same action class.

4.1.1. WEIZMANN DATASET

During initial test setups and the process of algorithm implementation the Weizmann dataset was used as given in [37]. This dataset was useful in terms of its steady background which enabled the validation of feature selection by inspection.

Each video clip in Weizmann dataset included single actor and single event. The events were periodic actions except one. The event classes were ‘running’, ‘walking’, ‘jumping jack (jack)’, ‘jumping forward on two legs’, ‘jumping forward on one leg’, ‘jumping in place on two legs (Pjump)’, ‘galloping-sideways’, ‘waving-two-hands’, ‘waving-one-hand’, ‘bending’. Each of 10 action types was performed once by 9 actors creating a total of 90 video clips. In the tests, the first 4 actor videos were taken as training videos and 5 actor videos were the test sets. A sample from each action is shown in Figure 4 and Figure 5.

In terms of video quality, Weizmann dataset is appropriate for vision purposes. Each video has resolution of 180x144 where the actor is about the half size of the frame height. The camera is stable and there is no background motion. A difficulty of the dataset concerns the similarity between some action types such as ‘jumping forward on two legs’ and ‘jumping forward in one leg’.

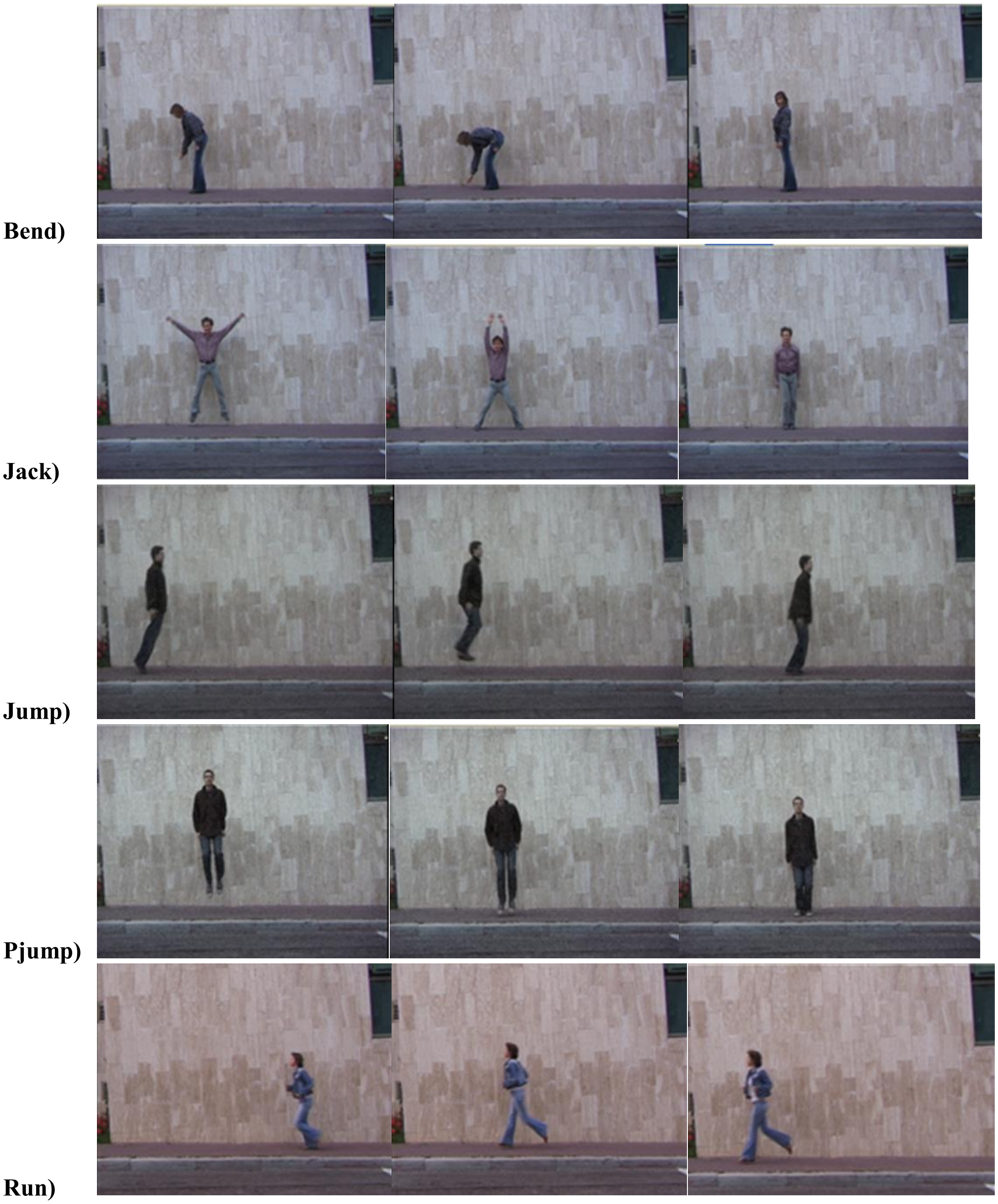
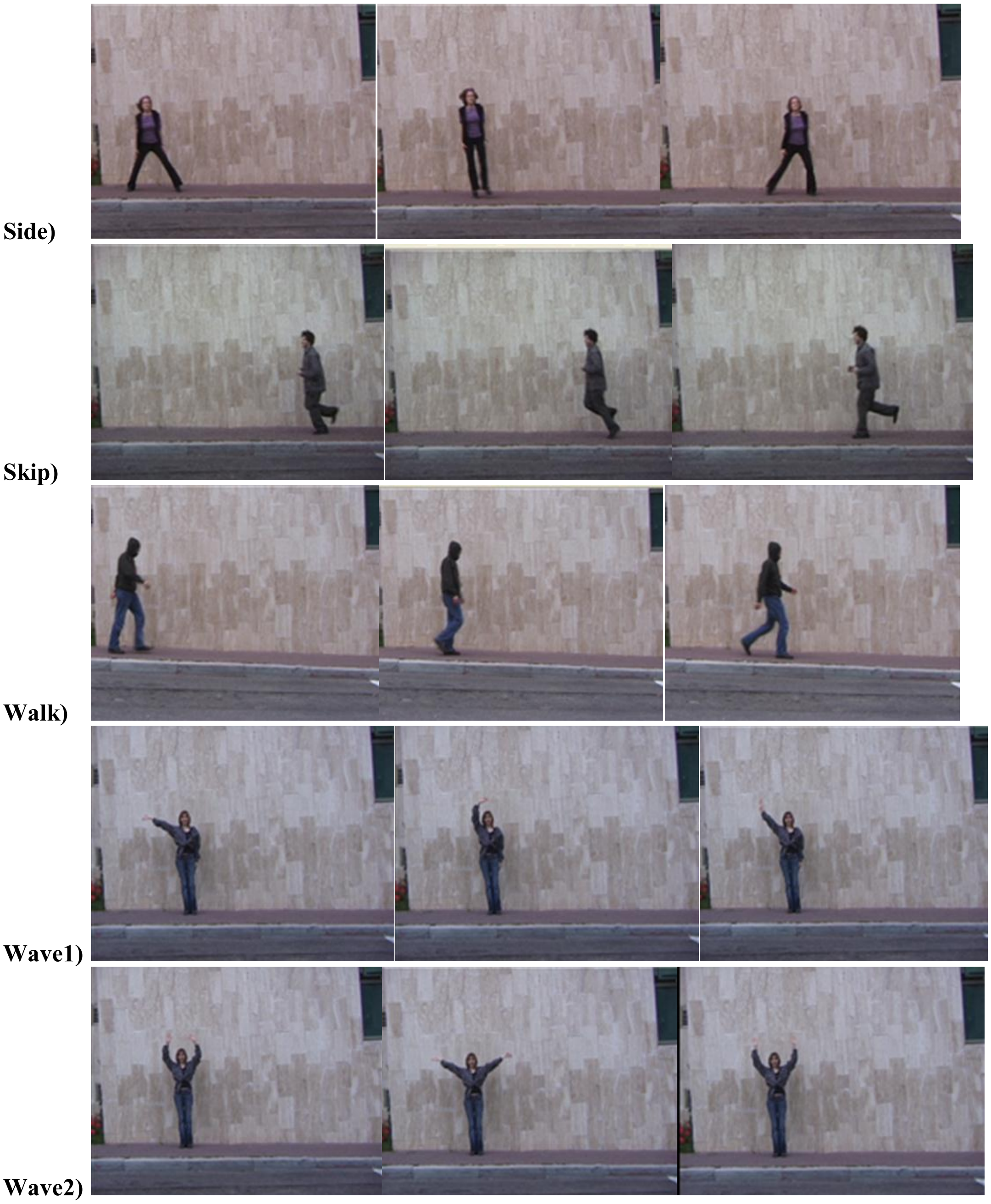


Figure 4: Weizmann Dataset samples – 1.



Side)

Skip)

Walk)

Wave1)

Wave2)

Figure 5: Weizmann Dataset samples – 2. Approximately a 90% scale is used.

4.1.2. KTH DATASET

Kth dataset which is given in [3] is almost the most popular dataset for classification of short video clips for activity recognition research. It is a challenging dataset with small background motion and environmental artifacts.

As in Weizmann dataset, each video clip in Kth dataset consists of a single actor and single event. The events are all periodic actions of; ‘boxing’, ‘handclapping’, ‘hand waving’, ‘jogging’, ‘walking’ and ‘boxing’. The sample frames for the actions are given in Figure 7. Six actions are performed by 25 actors several times in various conditions. The complete dataset contains 2,391 videos [4].

In the implementation for the current study, a setup similar to [34] was used, but for memory considerations, only 150 clips were used which is given as an outdoor set in [3]. Subjects 2, 3, 5, 6, 7, 8, 9, 10 and 22 were chosen as the training set, and the other 16 subjects were allocated as test set. Each subject performs 6 actions once.

In Kth dataset, the video clips were 160x120 in resolution and actors were about 90 pixels high. Videos have a slight background motion which made feature extraction process difficult. In some actions, the actors sometimes passed through the scene and returned after a while from the inverse direction (Figure 7). This makes the distance feature relations unrelated in Ryoo’s feature relations. A shadow also appears in some video clips which is stronger than the actor himself (Figure 6).

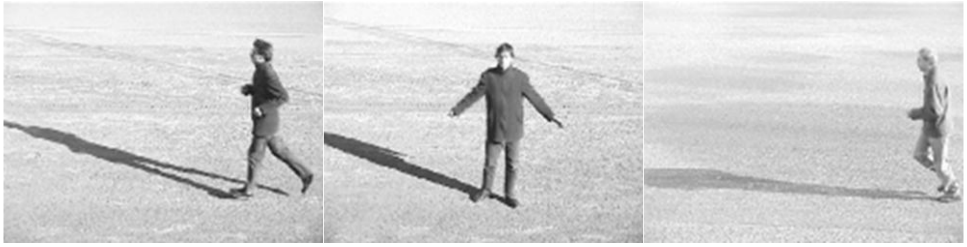


Figure 6: Shadows in KTH dataset

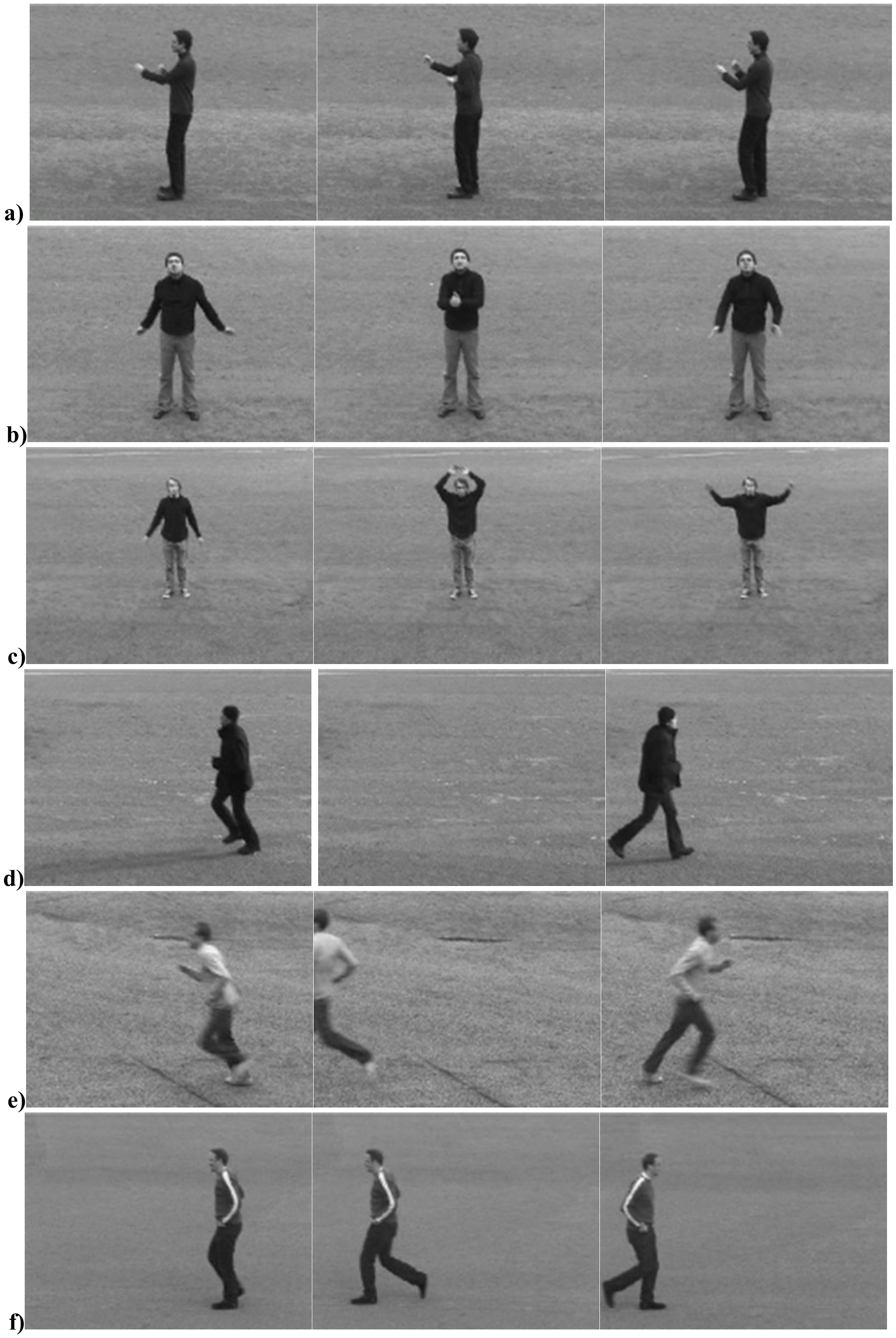


Figure 7: Kth Dataset samples. a) boxing, b)handclapping c)handwaving d)jogging e)running f)walking. Drawn with original scale. In c,d and e; subject leaves the scene and returns.

4.1.3. Non-Periodic Action Dataset

In the original work for feature relations [5], a new dataset was proposed for the classification of videos. Apart from the other datasets which includes periodic actions, this set includes non-periodic actions.

Although the complete dataset also includes large video sequences which contain several actors performing several non-periodic actions consecutively, this dataset was used as single action clips which are also presented in the original work and are similar to the setups in other datasets. In the sets in the current study, each action was performed once by one or two people depending on the activity type. The activities were; ‘handshaking’, ‘hugging’, ‘kicking’, ‘pointing’, ‘punching’ and ‘pushing’. Each of the activity classes were performed 20 times making a total of 120 videos which were divided equally between the test and training sets as in [5].

In this dataset, the resolutions vary since they are patches of clips which were extracted from high resolution videos. The actors were about 225 pixels in the original video and this was scaled them by 0.5 for memory considerations. After the scaling the size of the actors was in keeping with that of the other datasets so the same parameters were used in analyses.

There was no background motion in video clips however, in test set, there were foreign actors passing by in the background and there were actor, clothing and environment differences as shown in Figure 8. The most challenging part of this dataset was its activity types which were non-periodic and performed only once. This prevents the creation of statistically sufficient histograms during training.

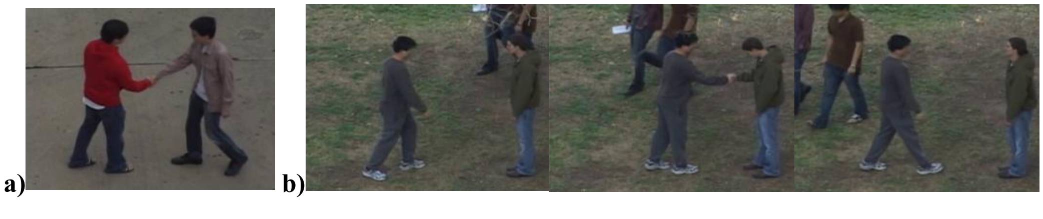


Figure 8: Challenges in Non-Periodic Action Dataset. There are actors passing by in the background, environment and clothing differences in the test set. a) Sample from training handshaking video, b) Sample from test handshaking video.

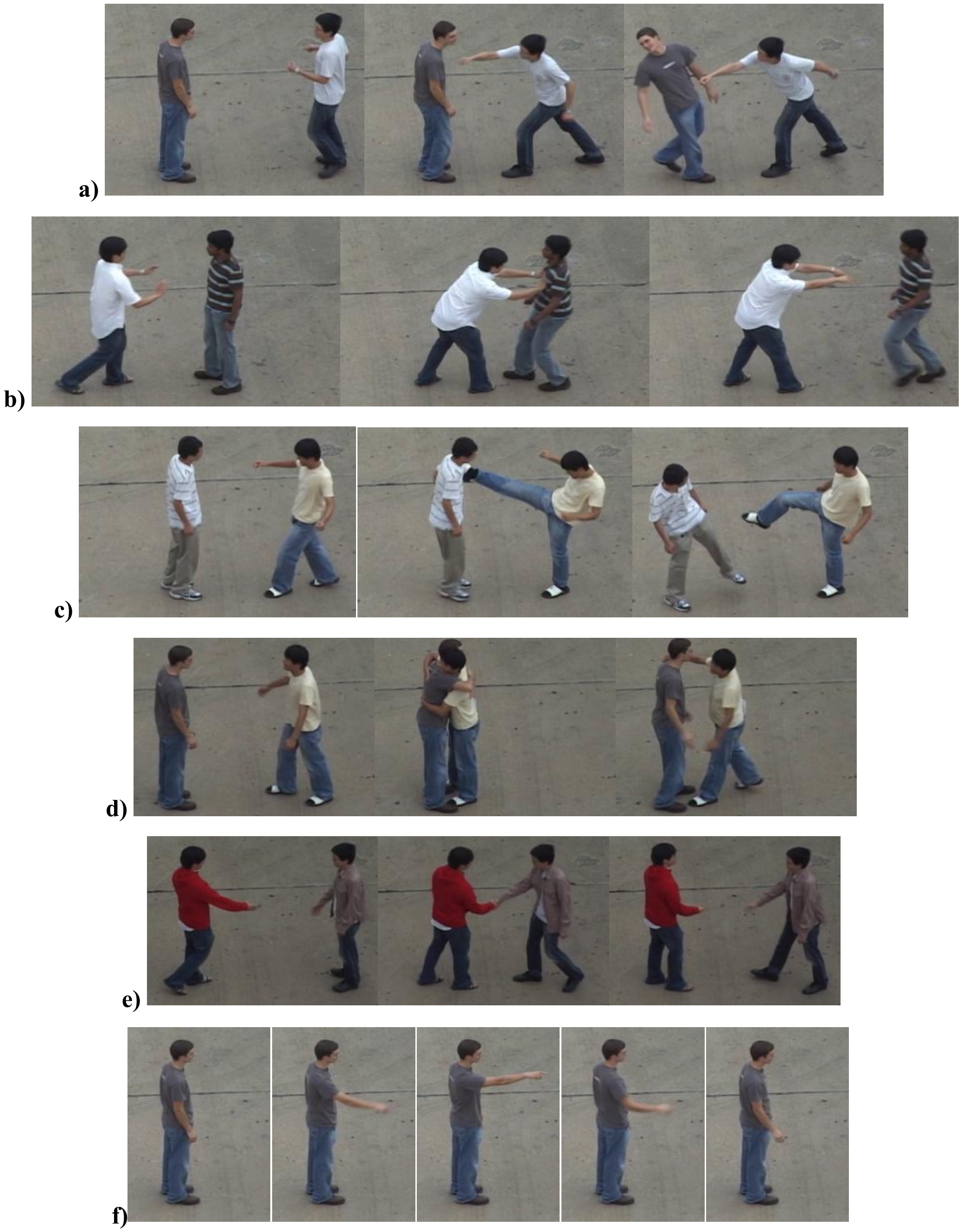


Figure 9: Non-Periodic Action Dataset samples with 35% scale. a) Punching, b) Pushing, c) Kicking, d) Hugging, e) Handshaking, f) Pointing.

4.2. Setups and Parameters

In the experimental setups, first the cuboid feature histogram method was used as proposed in [6]. This method was also used as a base for the other methods. The event descriptor and the event matching methods presented in sections 3.4 and 3.5 were implemented by making modifications on feature histogram method. Finally, we implemented our extensions which are proposed in 3.4.2.

Implementation was in Matlab environment. Technological support was Piotr Dollàr's image processing toolbox (version 1.3) which is available from his website[3] and Piotr Dollàr's cuboids toolbox which is available in his website[4] on request.

4.2.1. Feature Histogram Method

The Feature Histogram setup in the current study was based on cuboid feature histograms proposed in the original work [6]. The feature extraction process of this method is explained in sections 3.1, 3.2 and 3.3 and the classification process is explained in 3.5. In contrast to the other experimental setups detailed in this section, this setup uses feature histograms instead of feature relation histograms for action description. In the implementation, Piotr Dollàr's cuboid toolbox demo was configured to work with the datasets for the current study. Furthermore, the cuboid toolbox was modified in line with the memory requirements of the experimental setups.

Parameters

For cuboid method, if possible the original implementation parameters were maintained in toolbox demo. In the interest point selection process, the scale parameters were $\sigma = 2$ and $\tau = 3$. The 'periodic' type feature detection was used. In the toolbox, another option for this parameter was Harris3D which was proposed in [26]. A sample response function can be seen in Figure 3b.

[3] http://vision.ucsd.edu/~pDollàr/toolbox/doc/

[4] http://vision.ucsd.edu/~pDollàr/research.html#BehaviorRecognitionAnimalBehavior

During the feature selection, the number of interest points was not limited. The cuboid radius was 3 times multiple of the scale parameters as proposed in [6]. This made a 13x13x19 sized cuboid for each feature. For selection of the maxima points, the minimum distance between maxima was defined as 1 x 1 x 2 pixels. This takes in almost all local maximas in response function. A sample set of features is shown in Figure 3d.

The extracted cuboid features were converted into feature descriptors. Gradient values were used for the feature descriptor with no histogram. Two different 3D Gaussian smoothing was applied to the gradient values in order to cover the different scale information; $\sigma = 1, \tau = 0.5$ and $\sigma = 2, \tau = 0.5$.

PCA was used to shrink the descriptor size and increase the effect of common patterns. During implementation of PCA, a random sampling among features was used to obtain a set that was sufficient for the extraction of the PCA transformation matrix. 20 cuboids were selected from each clip.

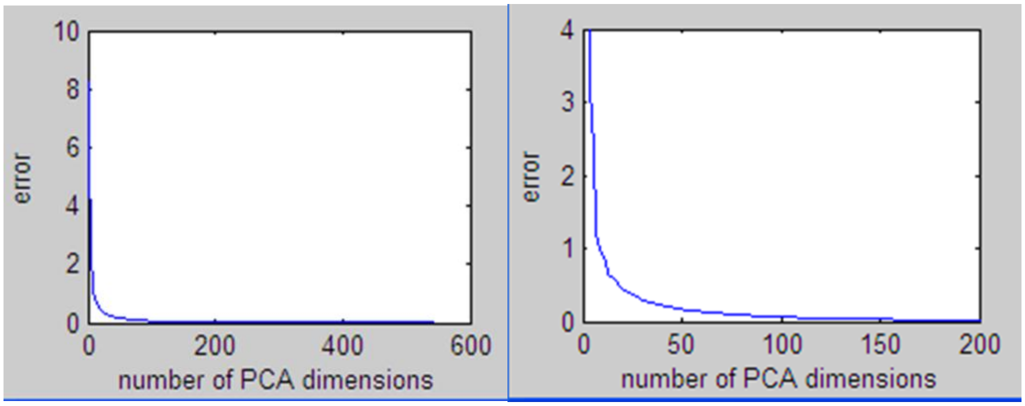


Figure 10: PCA error according to number of dimensions. (Same graph in different scales)

Figure 10 shows an analysis of the error rate for each level of the PCA dimensions. For the number of dimensions chosen, different numbers of dimensions from 20 to 100 were calculated however, there was little difference in the results. In the experiments, 100 dimensions were used as in the original implementation. After applying PCA, the features were clustered using k-means with 50 clusters.

The classification of the event descriptors was undertaken with a knn classifier with k=1.

4.2.2. SPATIO-TEMPORAL RELATION HISTOGRAM

This implementation was undertaken to compare the performance of Ryoo's spatio-temporal relation histogram method as explained in 3.4.1, with the histogram of features used in the feature histogram method. In the implementation in the current study, although the cuboid features were used, feature relation histograms were utilized rather than feature histograms. If possible, relations as in [5] were implemented, but the feature types were still as in the cuboids method. The parameters could also vary from original implementation since they are not explicitly given in the relevant article.

The temporal relation types in the histogram used in the current study were; '*equals', 'before', 'overlaps-before'*. The other types of relations explained in 3.4.1 are not valid since the features were fixed in size. The spatial relations in the proposed histogram were; *near, x-near, y-near*, and *far*. There were total of 7 features and for each 2 features in the activity, there were 2 relations.

In terms of the distance thresholds for the spatial features, several thresholds proportional to cuboid diameter were tested. The results from the current study were based on the distance threshold equal to the cuboid diameter.

The relation histogram was 50x50x7 histogram since there are 50 descriptor clusters and 7 relations.

4.2.3. SPATIO-TEMPORAL RELATION MATCH

In 3.5.1, a method for human activity detection is presented as proposed in [5] in which they evaluated detection ability of the method using trained thresholds for detection. In the current study, this method is evaluated in terms of its classification ability in comparison with other methods.

In the setup, as for the action description, the same configuration was used as explained in 4.2.2 above, but for the knn classification, spatio-temporal relation match distance criteria was used rather than the Chi-squared distance as the distance criteria of the knn classifier. This method is explained in section 3.5.1 and again, knn classifier is used with k=1.

4.2.4. Spatio-Temporal Feature Relation Histogram Using 3D Relations

For the 4th setup, the proposed extension was implemented in the feature relation histogram method. In this setup, the same configuration as in section 4.2.2 was used except that a new feature relation set was used as explained in section 3.4.2.

For the spatial relations 2 different sets were used. One set is the Cartesian product of *x-near,* x-*far* and *y-near* and *y-far* which is the set used in previous experiments. The other set is the Cartesian product of *x-near, x-far* and *y-near, y-above, y-below*. The cuboid spatial dimension was used as the near threshold. Feature pairs with distance 5 times the cuboid dimension in x or y were considered irrelevant and pairs were not counted in histogram.

For the temporal relations; *equals*, *overlaps, before* were used. Feature pairs with distance 2,5 times the cuboid temporal dimension were considered irrelevant and pairs were not counted in the histogram.

When the first spatial relation set was used, there were 4 spatial and 3 temporal relations; the size of the 3D relation set was 12. When the second spatial relation set was used there were 6 spatial and 3 temporal relations which make 18 3D relations.

4.3. Results and Discussion

Since it was the original implementation, the performance reference in the experiments was the feature histogram method explained in 4.2.1. In the experiments concerning the feature histogram method the success rates obtained were; 80% for the Weizmann dataset, 85% for the Kth dataset and 58% for the non-periodic dataset. This method achieved an 81.2% [5] and 87.7% [34] success rates in different implementations and training configurations of KTH dataset. Table 1 and Table 2 show the classification success rates in the first rows and the classification results are shown in Figure 11a, Figure 12a and Figure 13a.

In the second experimental setup, the feature relation match method was used. The tests on the relation match method showed that the relation match was not as successful as the feature histogram method for classification. The results can be seen in 2nd row of Table 1 and Figure 11b. A sliding window search with Chi-squared distance could be more successful in a detection process. As can be observe from the wave1 and wave2 actions in Figure 11, one action can reside in another. This is an important reason why there are false matches in the feature relation method.

Table 1: Comparison of the Relation Match Method on the Weizmann dataset.

Histogram type	Classifier	Success
Feature histogram	Knn+Chi-squared	0.83
Relation histogram.	Relation match	0.62

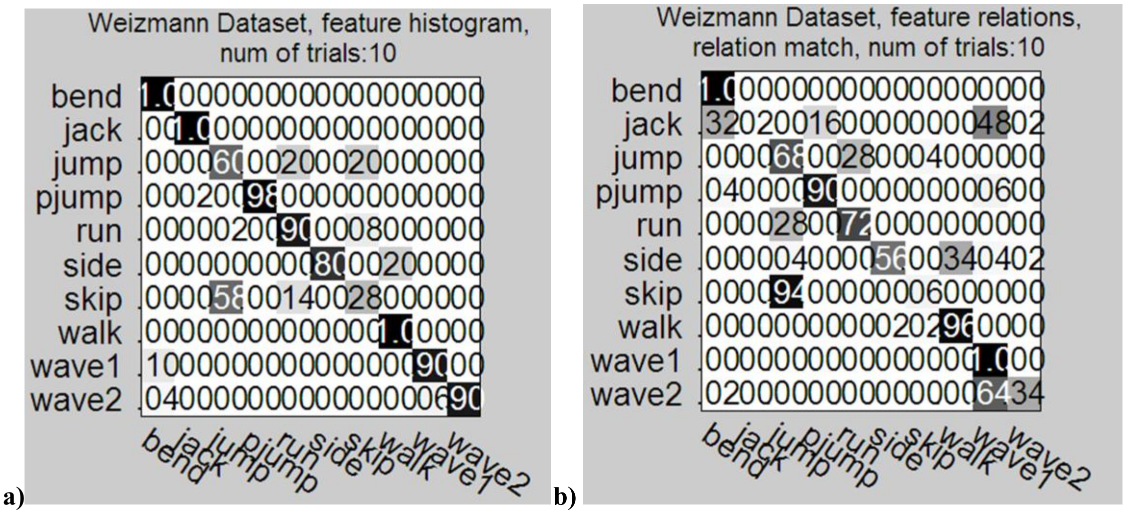


Figure 11: Classification results from Weizmann dataset. a) feature histogram method b) feature relation histogram match method.

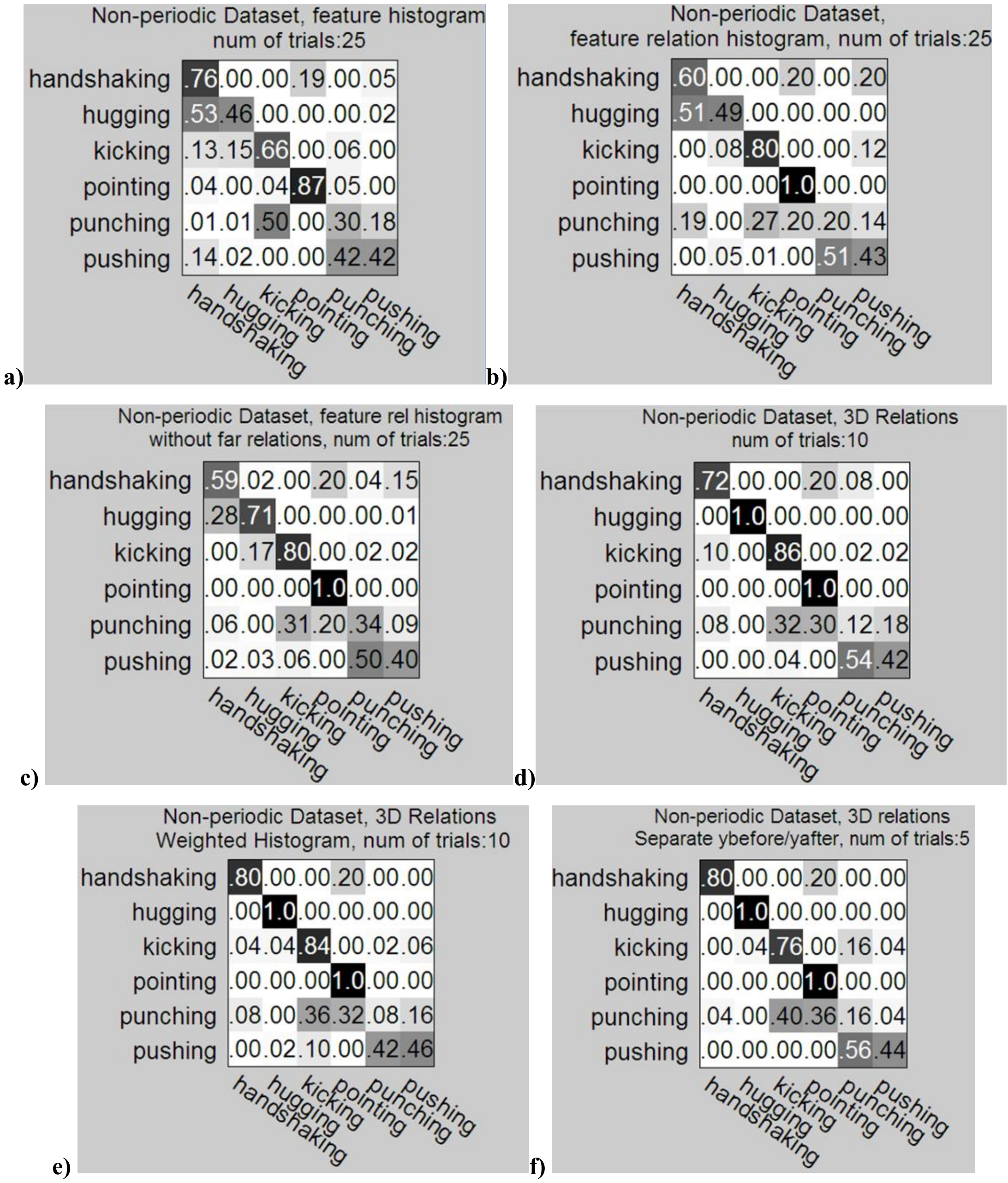
In the remainder of the experiments, a Chi-squared distance was chosen as the distance measure of the knn classifier. In addition the Kth dataset was selected for periodic actions since it is more populated.

In the third experimental setup, feature relation histograms were used with the same classifier of feature histogram method. For this method, slightly better results were obtained. The results are shown in 2nd row of Table 2, Figure 12b and Figure 13b for the Kth and the non-periodic datasets. About a 2% improvement on classification was observed in comparison with the feature histogram method.

Table 2: Comparison of histogram types; classification success on the Kth and Non-periodic datasets. The knn classifier is used.

	Datasets	
	KTH	Non-periodic
Feature Histogram	0.85	0.58
2D Spatio-temporal relations	0.87	0.59
2D Spatio-temporal relations without far relations	0.88	0.64
3D Spatio-temporal relations	0.87	0.69
3D Spatio-temporal relations with weighted histogram	0.87	0.70
3D Spatio-temporal relations. Separate above/below in y axis	0.87	0.70

Although there was a 2% improvement, higher results were observed according to the literature. As an example, a 3.4% increase from 87.7% to 91.1% exists between implementations on the same subset of Kth in [34] and [5]. Feature relations are also expected to perform well especially on non-periodic actions. The reasons for different results were difference in subsets of datasets and difference in feature extraction parameters.



a) Non-periodic Dataset, feature histogram num of trials:25

	handshaking	hugging	kicking	pointing	punching	pushing
handshaking	.76	.00	.00	.19	.00	.05
hugging	.53	.46	.00	.00	.00	.02
kicking	.13	.15	.66	.00	.06	.00
pointing	.04	.00	.04	.87	.05	.00
punching	.01	.01	.50	.00	.30	.18
pushing	.14	.02	.00	.00	.42	.42

b) Non-periodic Dataset, feature relation histogram, num of trials:25

	handshaking	hugging	kicking	pointing	punching	pushing
handshaking	.60	.00	.00	.20	.00	.20
hugging	.51	.49	.00	.00	.00	.00
kicking	.00	.08	.80	.00	.00	.12
pointing	.00	.00	.00	1.0	.00	.00
punching	.19	.00	.27	.20	.20	.14
pushing	.00	.05	.01	.00	.51	.43

c) Non-periodic Dataset, feature rel histogram without far relations, num of trials:25

	handshaking	hugging	kicking	pointing	punching	pushing
handshaking	.59	.02	.00	.20	.04	.15
hugging	.28	.71	.00	.00	.00	.01
kicking	.00	.17	.80	.00	.02	.02
pointing	.00	.00	.00	1.0	.00	.00
punching	.06	.00	.31	.20	.34	.09
pushing	.02	.03	.06	.00	.50	.40

d) Non-periodic Dataset, 3D Relations num of trials:10

	handshaking	hugging	kicking	pointing	punching	pushing
handshaking	.72	.00	.00	.20	.08	.00
hugging	.00	1.0	.00	.00	.00	.00
kicking	.10	.00	.86	.00	.02	.02
pointing	.00	.00	.00	1.0	.00	.00
punching	.08	.00	.32	.30	.12	.18
pushing	.00	.00	.04	.00	.54	.42

e) Non-periodic Dataset, 3D Relations Weighted Histogram, num of trials:10

	handshaking	hugging	kicking	pointing	punching	pushing
handshaking	.80	.00	.00	.20	.00	.00
hugging	.00	1.0	.00	.00	.00	.00
kicking	.04	.04	.84	.00	.02	.06
pointing	.00	.00	.00	1.0	.00	.00
punching	.08	.00	.36	.32	.08	.16
pushing	.00	.02	.10	.00	.42	.46

f) Non-periodic Dataset, 3D relations Separate ybefore/yafter, num of trials:5

	handshaking	hugging	kicking	pointing	punching	pushing
handshaking	.80	.00	.00	.20	.00	.00
hugging	.00	1.0	.00	.00	.00	.00
kicking	.00	.04	.76	.00	.16	.04
pointing	.00	.00	.00	1.0	.00	.00
punching	.04	.00	.40	.36	.16	.04
pushing	.00	.00	.00	.00	.56	.44

Figure 12: Classification results from non-periodic action dataset. a) Feature histogram method b) Feature relation histogram method with Ryoo's relations. c) Feature relation histogram method without far relations d) Feature relation histogram method with 3D relations. e) Feature relation histogram method with 3D relations and weighted histogram. f) Feature relation histogram method with separate above/below relations

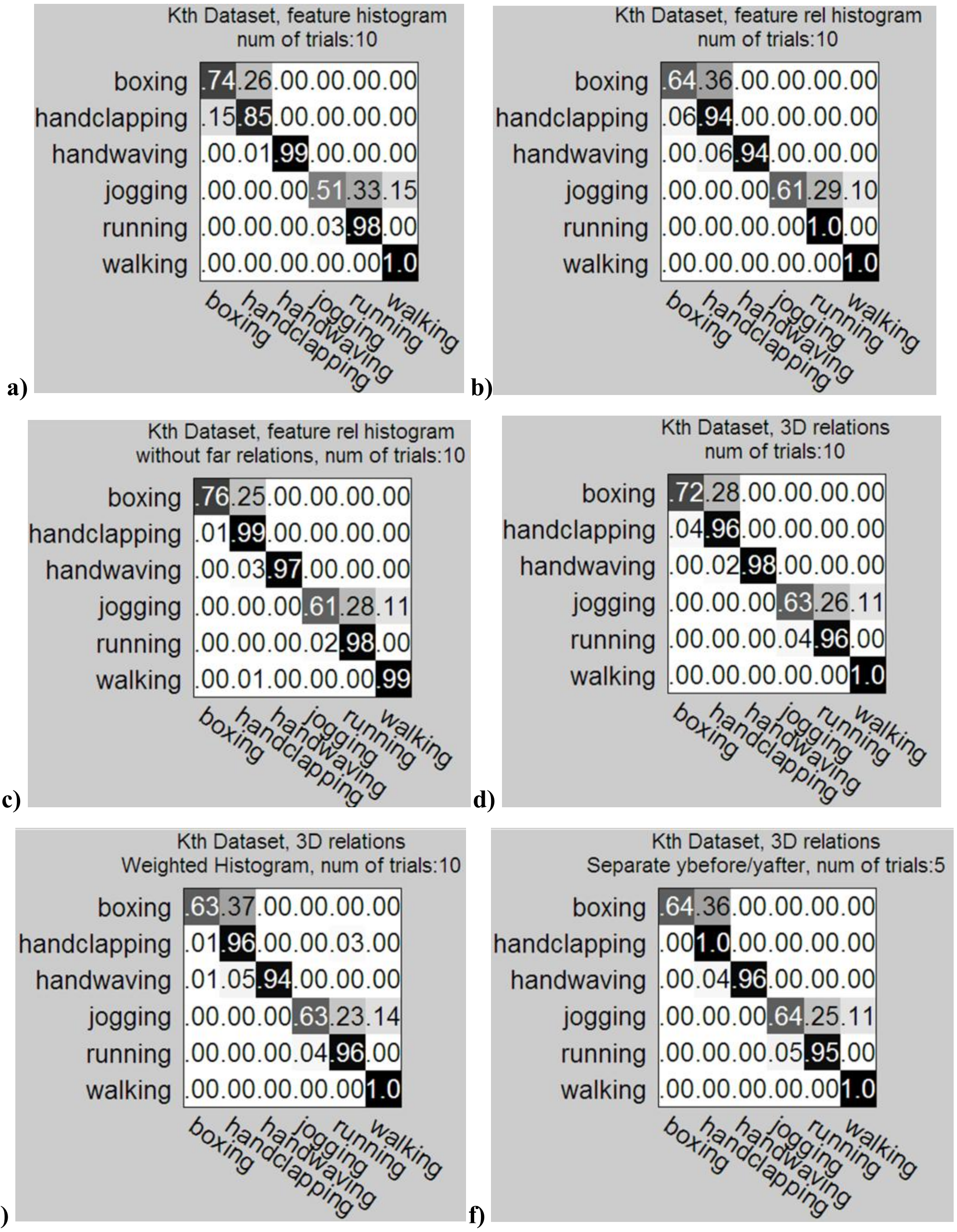


Figure 13: Classification results from Kth action dataset. a) Feature histogram method b) Feature relation histogram method with Ryoo's relations. c) Feature relation histogram method without far relations d) Feature relation histogram method with 3D relations. e) Feature relation histogram method with 3D relations and weighted histogram. f) Feature relation histogram method with separate above/below relations

In contrast, the setup in the current study uses a better criterion for a comparison. It uses the same feature extraction procedure for the feature descriptors. Even though Ryoo used cuboid features in his original study [5], his cuboid feature method has different parameters. Some known differences are: he used intensity cuboids, where gradient cuboids were used in the current study; he used 500 clusters, where we 50 were used in the current study.

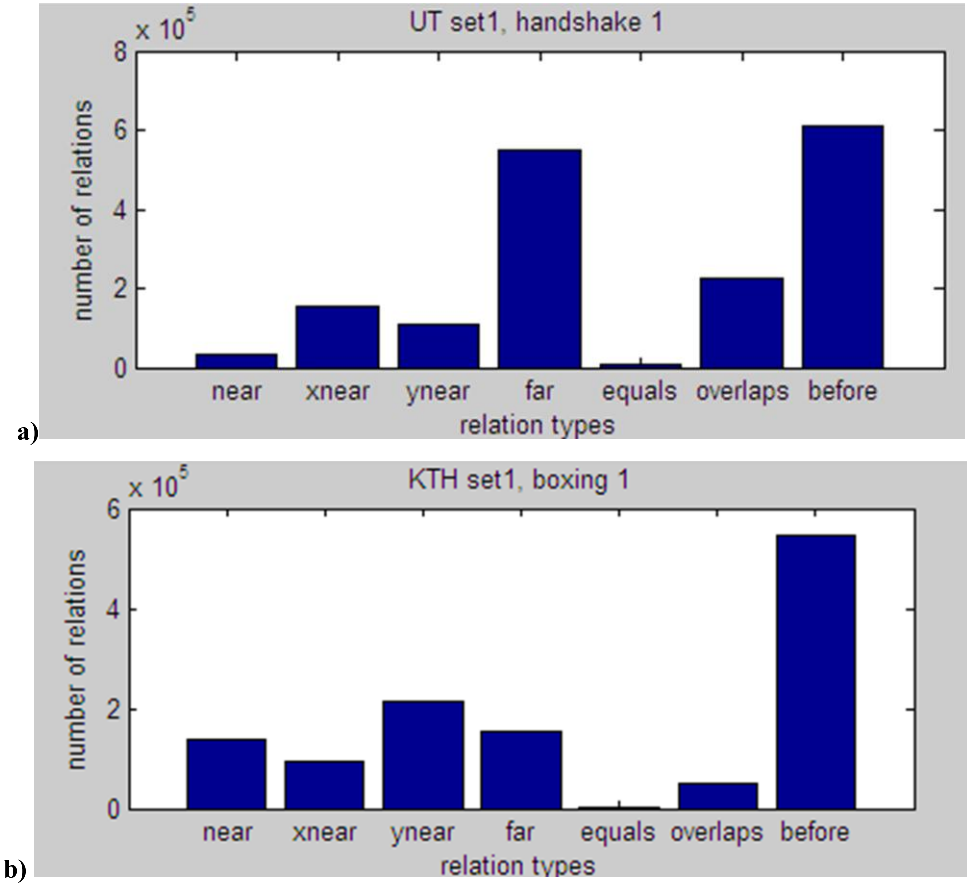


Figure 14: Distribution of Ryoo's feature relation types. a) Sample from Non-Periodic Dataset, b)Sample from KTH Dataset

In the relation histogram classification results, an expected reason for the low performance was the gap between histogram bins. An analysis of histogram bins was undertaken to detect which relations affect our results. The current study on the relation histogram shows that number of far relations is much higher than the number of close relations. An example of relation distributions can be observed in Figure 14. This shows that these results depend on far relations much more than close relations. To observe the effect of far relations in the experiments, *far* was removed from the spatial histogram and *before* from the temporal histogram. Ryoo's relation histogram without far relations shows 6% better performance on the non-periodic dataset when compared to the feature histogram method. The results are shown in the 3rd row of Table 2. From this experiment, it can be suggested that feature relations are better if they are used in weighted histograms. Close relations in terms of spatial and temporal distance need higher weights because they have a superior ability to represent an event. The relevance of features decrease when there is a large distance between them.

Another setup used in the experiments in the current study was for the 3D feature relation histogram method. The experimental setup for this method is explained in section 4.2.4 and the results are given in the 4th row of Table 2 and comparison charts are given in Figure 12d and Figure 13d. A 12% performance increase on non-periodic datasets was observed in comparison to feature histogram method. This shows that 3D relation histogram is more successful in representing non-periodic activities than any other method that was tested above.

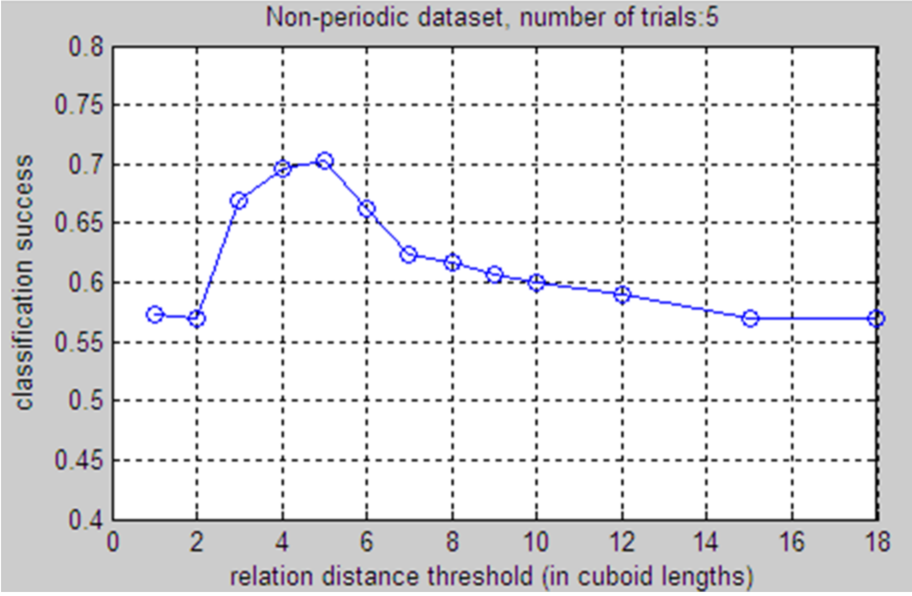


Figure 15: Effect of relation distance threshold

On 3D Relations method, effect of removal of far relations above some threshold is studied. Figure 15 shows classification success rates for a range of distance thresholds in terms of cuboid sizes. Experiments show that omitting relations above some distance threshold gives better results in relation histogram method. From these experiments, the optimal threshold is measured as 5 cuboid lengths in any direction.

The proposed new type of histogram was also tested with weights. These weights were calculated as a rough normalization of the bins and are obtained from the observations of a few histograms as shown in Table 3. The result from the setup is given in the 5th row of Table 2 and in Figure 12d. In comparison to the previous experiment without weights there was no great difference. The reason is, that the feature relations in the current study did not include far relations above a certain threshold, thus there was little difference among the histogram bins as found in Ryoo's histograms.

Table 3: Weights used for weighted 3D Relation histogram

	Near	x-near	y-near	Far
Equals	1	0.5	0.5	0.25
Overlaps	1	0.5	0.5	0.25
Before	1	0.5	0.5	0.25

Another experimental setup in our study uses directional relations in y axis in 3D relations. This set is explained as the second set in 4.2.4. The comparison results for this set is given in 6th row of Table 2 and comparison charts are given in Figure 12f and Figure 13f. This relation set also showed a close performance result to the previous feature relation set.

The experiments in the current study on different datasets show that the improvement of the feature relation histogram is particularly relevant for non-periodic datasets as also suggested in [5]. For periodic datasets, the feature histograms also showed a close performance.

CHAPTER 5

CONCLUSION

In this thesis, the feature histogram, feature relation histogram and feature relation histogram match methods are applied to human activity classification. For the feature relation histogram method, new types of histogram types are proposed and applied. Cuboid features are used as the feature type in all the experimental setups. The classification results show that feature relation histograms are more successful in representing the activity when compared to feature histograms. This performance is even greater with the proposed 3D relation set.

The classification results presented in this thesis show an adequate evaluation of the algorithms. Ryoo [5], achieved a success rate which is higher than the results from the current study however, his comparison of feature histogram methods was made by referencing previous works. That comparison did not use the same feature extraction process as the methods that are referenced thus the results may not be a fair comparison of the relation histogram with the feature histogram. In the current work, the same feature extraction process was used for all methods and the comparison was undertaken using the same datasets.

The results presented in this thesis show that the proposed feature relation histogram performs better than feature histograms for activity classification. When method of feature relation histogram was implemented in a manner similar to the original work, it gave an improvement of 1%-2% for Kth and Non-periodic datasets. Although this result is not sufficient enough to claim an improvement, the results for non-periodic dataset using modified versions of those relation histograms or new types of feature relation histograms proposed in this study shows about 12% improvement. This proves that the proposed feature relation histogram method is superior.

This study demonstrates the classification accuracy of the feature relation match method [5] which is proposed for event detection. The classification performance of this method is evaluated. Instead of looking for events over a certain threshold for detection, the best match is used over the class of events for classification. This relation match method is used as a replacement for the knn classifier used in other methods.

The results show that the classification performance of knn classifier is much higher than the relation match method. This difference in classification performance is also a criterion for the success of the relation match method. The results prove that false detection rate of relation match can be improved by using classifiers on detection results.

Another contribution of this study is the proposed new set of relations for the relation histogram. Ryoo's relation histogram [5] holds spatial and temporal relations in different bins, however, close spatial relations between features do not need to have an important relation if they are far apart in temporal axis. The proposed feature relation histogram uses a new set of relations consisting of 3D relations. The classification results from the current study show that the new method is able present activity better than the previous types of relation histogram.

In fact, computational cost of the proposed scheme was not considered during the research. From the size of feature representation histograms, feature relation histogram methods will require more computational time and memory for comparisons. Depending on the number of 3D relation, the proposed feature histograms may require even more time and memory. On the other hand, it was observed that the feature descriptor extraction process was the most time consuming operation during the experiments thus those methods are also open to improvements in terms of computational cost.

This study presents a comparison and suggests an improvement for feature relation histogram methods. The results are intended to be a guide for further developments in the area of human activity recognition. The feature relation histogram method for activity representation is a promising application in the study of lower level activity recognition. Better representation of lower level activities will also lead developments on hierarchical recognition methods which are able to resolve more complex activity types.

BIBLIOGRAPHY

[1] Vigilant Video, [Online]. Available: http://www.vigilantvideo.com/smartz.htm. [Accessed 26 06 2012].

[2] Z. Wu and R. Radke, "Real-time airport security checkpoint surveillance using a camera network," in *CVPRW*, 2011.

[3] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach," in *ICPR*, 2004.

[4] J. Aggarwal and M. S. Ryoo, "Human Activity Analysis: A Review," *ACM Computing Surveys,* vol. 43, no. 3, pp. 16:0-43, April 2011.

[5] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *ICCV*, Kyoto, Japan, 2009.

[6] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Beijing, 2005.

[7] M. Perkowitz, M. Philipose and K. Fishkin, "Mining models of human activities from the web," in *Proceedings of the 13th international conference on World Wide Web*, New York C., 2004.

[8] H. Ning, T. Han, D. Walther, M. Liu and T. Huang, "Hierarchical Space-Time Model Enabling Efficient Search for Human Actions," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 19, no. 6, pp. 808-820, 2009.

[9] V. Pavlovic, R. Sharma and T. Huang, "Visual interpretation of hand gestures for human-computer interaction: a review," *IEEE Transactions Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 677-695, 1997.

[10] T. Starner and A. Pentland, "Real-time American Sign Language recognition from video using hidden Markov models," in *Proceedings., International Symposium on Computer Vision*, 1995.

[11] I. Laptev, M. Marsza, C. Schmid and B. Rozenfeld, "Learning Realistic Human Actions from Movies," in *CVPR*, 2008.

[12] T. B. Moeslund and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture," *Computer Vision and Image Understanding,* vol. 81, no. 3, pp. 231-268, 2001.

[13] T. B. Moeslund, A. Hilton and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding,* vol. 104, no. 2-3, pp. 90-126, 2006.

[14] B. Kepenekçi, Human Activity Recognition by Gate Analysis, a thesis submitted to Grad. S. of Nat. and App. Sci., Ankara: METU, 2011.

[15] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257- 267, 2001.

[16] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *CVPR*, 2005.

[17] Y. Ke, R. Sukthankar and M. Hebert, "Spatio-temporal Shape and Flow Correlation for Action Recognition," in *CVPR*, 2007.

[18] Y. Sheikh, M. Sheikh and M. Shah, "Exploring the space of a human action," in *ICCV*, 2005.

[19] T. Darrell and A. Pentland, "Space-time gestures," in *CVPR*, 1993.

[20] D. Gavrila and L. Davis, "Towards 3-D model based tracking and recognition of human movement: a multi-view approach," in *Int. Workshop on Face and Gesture Recognition*, Zurich, 1995.

[21] R. Lublinerman, N. Ozay, D. Zarpalas and O. Camps, "Activity Recognition from Silhouettes using Linear Systems and Model (In)validation Techniques," in *ICPR*, 2006.

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

images using hidden Markov model," in *CVPR*, 1992.

[23] N. Oliver, B. Rosario and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 831 - 843, 2000.

[24] S. Park and J. K. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions," *Multimedia Systems,* vol. 10, no. 2, pp. 164-179, 2004.

[25] P. Natarajan and R. Nevatia, "Coupled Hidden Semi Markov Models for Activity Recognition," in *WMVC (IEEE Workshop on Motion and Video Computing)*, 2007.

[26] I. Laptev, "On Space-Time Interest Points," *International Journal of Computer Vision,* vol. 64, no. 2, pp. 107-123, 2005.

[27] P. Scovanner, S. Ali and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th international conference on Multimedia*, Augsburg, Germany, 2007.

[28] A. Klaser and M. Marszalek, "A spatio-temporal descriptor based on 3D-gradients," in *British Machine Vision Conference*, Leeds, 2008.

[29] C. Harris and M. Stephens, "A Combined Corner and Edge Detection," in *In Proceedings of The Fourth Alvey Vision Conference*, 1988.

[30] D. Lowe, "Object recognition from local scale-invariant features," in *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999.

[31] S.-F. Wong, T.-K. Kim and R. Cipolla, "Learning Motion Categories using both Semantic and Structural Information," in *CVPR*, 2007.

[32] J. Niebles, H. Wang and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *International Journal of Computer Vision,* vol. 79, no. 3, pp. 299-318, 2008.

[33] S. Savarese, A. DelPozo, J. Niebles and L. Fei-Fei, "Spatial-Temporal correlatons for unsupervised action classification," in *IEEE Workshop on Motion and video Computing*, 2008.

[34] H. Wang, M. M. Ullah, A. Klaser, I. Laptev and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision*

Conference, London, 2009.

[35] M. Bregonzio, S. Gong and T. Xiang, "Recognising action as clouds of space-time interest points," in *CVPR*, 2009.

[36] G. Bebis, "Support Vector Machines (SVM), Lecture Notes," Dep. of CSE at the U. of Nevada, [Online]. Available: http://www.cse.unr.edu/~bebis/MathMethods/SVM/lecture.pdf. [Accessed 28 06 2012].

[37] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes," in *ICCV*, 2005.