

# Human Action Recognition Using 3D Joint Information and Pyramidal HOOFD Features

by

Barış Can Üstündağ

Submitted to the Graduate School of Sabanci University  
in partial fulfillment of the requirements for the degree of  
Master of Science

Sabanci University

July, 2014

Human Action Recognition Using 3D Joint Information and  
Pyramidal HOOFD Features

APPROVED BY:

Prof. Dr. Mustafa Ünel  
(Thesis Advisor)

.....

Assoc. Prof. Dr. Kemalettin Erbatur

.....

Assoc. Prof. Dr. Erkay Savaş

.....

DATE OF APPROVAL:

.....

© Barış Can Üstündağ 2014  
All Rights Reserved

# Human Action Recognition Using 3D Joint Information and Pyramidal HOOFD Features

Bariş Can Üstündağ

ME, Master's Thesis, 2014

Thesis Supervisor: Prof. Dr. Mustafa Ünel

Keywords: Action Recognition, Classification, RGBD Images, Depth Data,  
HOOFD

## Abstract

With the recent release of low-cost depth acquisition devices, there is an increasing trend towards investigation of depth data in a number of important computer vision problems, such as detection, tracking and recognition. Much work has focused on human action recognition using depth data from Kinect type 3D cameras since depth data has proven to be more effective than 2D intensity images.

In this thesis, we develop a new method for recognizing human actions using depth data. It utilizes both skeletal joint information and optical flows computed from depth images. By drawing an analogy between depth and intensity images, 2D optical flows are calculated from depth images for the entire action instance. From the resulting optical flow vectors, patches are extracted around each joint location to learn local motion variations. These patches are grouped in terms of their joints and used to calculate a new feature called 'HOOFD' (Histogram of Oriented Optical Flows from Depth). In order to encode temporal variations, these HOOFD features are calculated in a pyramidal fashion. At each level of the pyramid, action instance is partitioned equally into two parts and each part is employed separately to form the histograms. Oriented optical flow histograms are utilized due to their invariance to scale and direction of motion. Naive Bayes and SVM classifiers are then trained using HOOFD features to recognize various human actions. We performed several experiments on publicly available databases and compared our approach with state-of-the-art methods. Results are quite promising and our approach outperforms some of the existing techniques.

# 3D Eklem Bilgisi ve Piramit HOOFD Özniteliğini Kullanarak İnsan Aktivitelerini Tanıma

Bariş Can Üstündağ

ME, Master Tezi, 2014

Tez Danışmanı: Prof. Dr. Mustafa Ünel

Anahtar Kelimeler: Aktivite Tanıma, Sınıflandırma, RGBD İmgeler,  
Derinlik Verisi, HOOFD

## Abstract

Düşük maliyetli derinlik yakalayan cihazların piyasaya sürülmesiyle tespit, takip ve tanıma gibi birçok önemli bilgisayarla görme probleminde derinlik verisinin kullanımı yükselen bir trend haline geldi. Kinect 3D kamerası kullanılarak insan aktivitelerini tanıma konusu üzerine de bir çok çalışma yapılmış ve bu bağlamda derinlik verisinin 2D imgelerden daha efektif olduğu kanıtlanmıştır.

Biz bu tezde derinlik verisinden insan aktivitelerini tanıma üzerine yeni bir yöntem geliştirdik. Bu yöntem hem 3D eklem bilgisini hem de derinlik imgelerinden hesaplanan optik akışı kullanmaktadır. Derinlik ve yoğunluk imgeleri arasında kurduğumuz bağıntı doğrultusunda derinlik imgelerinden 2D optik akış vektörleri bütün bir aktivite örneği süresince hesaplanmaktadır. Sonra, 3D eklem konumları baz alınarak bölgesel hareket değişimlerini öğrenebilmek için her bir eklem çevresinden optik akış vektörlerini içeren parçalar çıkartılmaktadır. Bu parçalar bulunduğu ekleme göre gruplanıp geliştirdiğimiz HOOFD (Histogram of Oriented Optical Flows from Depth) özniteliğini hesaplamakta kullanılmaktadır. Zamansal değişimleri de takip edebilmek için HOOFD öznitelikleri piramitsel bir yaklaşımla hesaplanmıştır. Piramidin her seviyesinde aktivite eşit iki bölüme ayrılıp her bölüm histogramları doldurabilmek için ayrı değerlendirilmiştir. Ölçek ve hareket yönü değişmezliği avantajlarından dolayı optik akış vektörlerinin yönelimlerinden oluşan histogramlar kullanılmıştır. Naive Bayes ve Destek Vektör Makinaları (DVM) sınıflandırıcıları HOOFD öznitelikleri kullanılarak eğitilmiş ve birbirinden farklı birçok aktiviteyi tanımak için kullanılmıştır.

Farklı veri kümeleri ile birçok deney yapılmış ve önerilen yöntem literatürdeki en gelişkin yöntemlerle karşılaştırılmıştır. Sonuçlar oldukça umut vericidir ve önerdiğimiz yöntem mevcut bazı tekniklerden daha iyi performans göstermektedir.

## Acknowledgements

First of all, I would like to express my sincere gratitude to my thesis advisor Prof. Dr. Mustafa Ünel for his supervision, endless encouragement and mental support. His wisdom and passion for computer vision and for life taught me a lot.

I am gratefully thanking my fellow colleagues Taygun Kekeç, Alper Yıldırım, Soner Ulun, Mehmet Ali Güney, Caner Şahin and all remaining CVR Research group members for hours of discussions, brainstormings and collaboration.

Finally, I would like to thank my family, my brother, my parents and my soulmate Irem for all their love and support throughout my life. I would not be able to accomplish anything without each and every single one of them.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	3
1.2	Thesis Contributions and Organization . . . . .	5
<b>2</b>	<b>Related Work</b>	<b>7</b>
2.1	Intensity based methods . . . . .	7
2.2	Depth map based methods . . . . .	10
2.3	Skeletal data based methods . . . . .	11
<b>3</b>	<b>Action Recognition using Depth Data</b>	<b>17</b>
3.1	Acquiring Depth Data . . . . .	17
3.2	Feature Extraction . . . . .	19
3.2.1	Joint Features . . . . .	20
3.2.2	Optical Flow from Depth Data . . . . .	23
3.3	Feature Representation . . . . .	24
3.3.1	Signal Warping . . . . .	24
3.3.2	Patch Extraction and HOOFD Features . . . . .	24
3.4	Classification Methods . . . . .	29
3.4.1	Naive Bayes Classifier . . . . .	29
3.4.2	Support Vector Machines . . . . .	29
<b>4</b>	<b>Experiments</b>	<b>31</b>
4.1	Datasets . . . . .	31
4.1.1	MSR Action3D Dataset . . . . .	31
4.1.2	MSR Action Pairs Dataset . . . . .	32
4.1.3	MSRC-12 Gesture Dataset . . . . .	33



4.2	Joint Features with Signal Warping . . . . .	35
4.3	Pyramidal HOOFD Features . . . . .	36
<b>5</b>	<b>Conclusion &amp; Future Work</b>	<b>43</b>

## List of Figures

1.1	Depth estimation techniques . . . . .	1
1.2	Images acquired from the Kinect sensor . . . . .	2
1.3	Sensors that acquires depth data . . . . .	3
1.4	Cameras and a CCTV station . . . . .	4
2.1	Extraction of cuboids from two different action instances, even though the posture of the mouse is quite different, extracted cuboid patches are similar [18] . . . . .	8
2.2	MHI and MEI notions proposed in [21] . . . . .	9
2.3	Generated 3D surface normals are illustrated in the work of [33] . . . . .	11
2.4	Illustration of the most informative joints during an action instance [40] . . . . .	13
3.1	Flow Chart of the Proposed Method . . . . .	17
3.2	Illustrating the cause of shadow . . . . .	19
3.3	Joint Features Illustration . . . . .	21
3.4	An illustration of joint angle calculation by defining vectors between each joint locations . . . . .	22
3.5	Mapping depth data to grayscale intensity image . . . . .	23
3.6	Randomly selected frames are discarded / replicated and inserted in a action sequence . . . . .	25
3.7	Overview of the proposed method . . . . .	28
3.8	SVM classifier returns maximum margin decision boundary (hyperplane) . . . . .	30
4.1	Depth image sequence examples from MSRAction3D dataset . . . . .	32
4.2	Depth image examples of MSR Action Pairs dataset . . . . .	33

4.3	Gestures and captured frames from gesture instances of MSRC-12 dataset [70] . . . . .	34
4.4	Confusion matrices of MSRC-12 dataset using (1:1) experimental settings . . . . .	37
4.5	Confusion matrix of different action sets under Cross Subject Test . . . . .	38
4.6	Visualization of the skeleton tracker Failure on bend action . .	39
4.7	Confusion Matrix of MSR Action Pairs dataset under Cross Subject Test . . . . .	41
4.8	Recognition results for comparing patch size . . . . .	42

## List of Tables

4.1	Actions of MSRAction3D are divided into 3 subsets (numbers in paranthesis represents the action annotations) . . . . .	32
4.2	Feature sets are generated in order to use on MSRC-12 Gesture dataset . . . . .	35
4.3	Recognition accuracies (%) Comparison of different tests for MSRC-12 Gesture dataset . . . . .	36
4.4	Recognition accuracies (%) Comparison of Cross Subject Test for MSR Action 3D . . . . .	36
4.5	Comparison of classification accuracy with state-of-the-art methods for MSRAction3D dataset . . . . .	40
4.6	Classification accuracy comparison for MSR Action Pairs dataset	40
4.7	Classification accuracy of our method at each pyramid level .	41

# Chapter I

## 1 Introduction

Computer vision is one of the most active and flourishing disciplines among today's research areas. Various solutions are proposed to the problems of detection, recognition and tracking objects. Most of them employ heuristic approaches that use 2D intensity images, even though we are living and interacting in a 3D world.

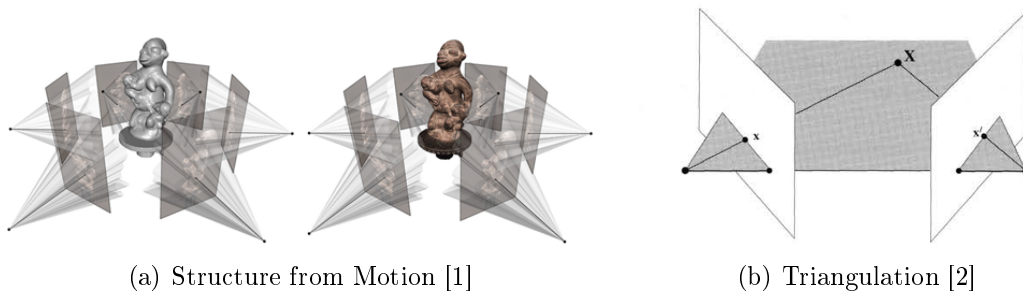


Figure 1.1: Depth estimation techniques

One of the most challenging tasks in computer vision is to estimate 3D depth data. For a computer it is impossible to understand the depth information from a single 2D image. Even though there are several estimation methods, e.g. structure from motion [1] and triangulation [2] that achieved promising results, they are not robust enough in certain real world scenarios. Another method for 3D depth estimation is using range sensors or motion capture systems.

Earlier range sensors were quite expensive and not easily accessible, and their application range was limited upto 6-7 meters.

Marker-based motion capture systems are also used to extract movement of the people in the 3D environment. However even today they are expensive and need a static working environment for the installation of the cameras.

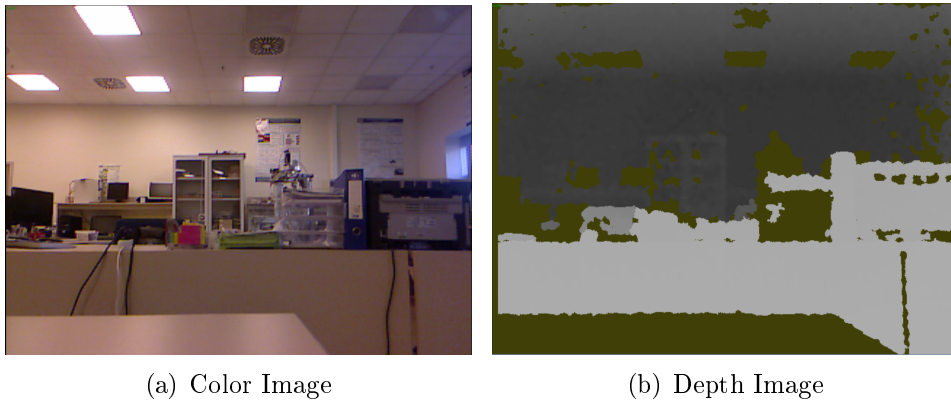


Figure 1.2: Images acquired from the Kinect sensor

With the release of Microsoft Kinect [3] and low-cost and relatively accurate other sensors such as ASUS Xtion Pro Live, it has become easier to capture depth information. Despite its initial purpose, which was supporting a gaming console for interactive gaming, it also took lots of attention by scientific authorities who are in the fields of robotics, health and medicine, education, and vision. Due to Kinect's real-time depth capture feature various computer vision problems can be solved using very low computational power.

Latest products that are spread to the market, e. g. Leap Motion, LG G3 smartphone and even the upcoming Google Tango prove that the depth data acquired directly from a sensor can be used in many important applications. In this thesis we used depth information to be able to recognize human



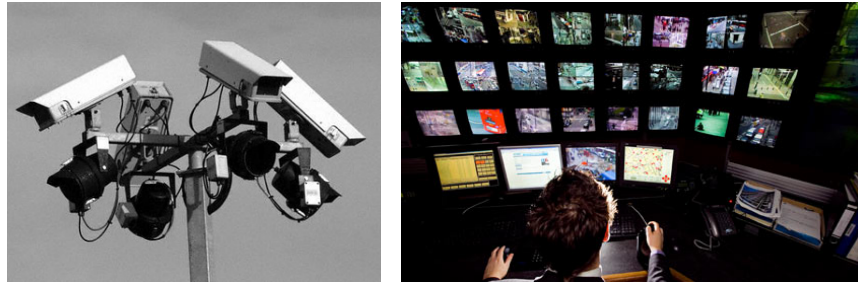
Figure 1.3: Sensors that acquire depth data

actions.

## 1.1 Motivation

Human action recognition is one of the most active areas in computer vision. Due to its importance in a number of real-world applications e.g. human-computer interaction, health-care, surveillance and smart-home applications, it maintains its significance among other research areas. One of those areas, human machine interaction, possess the highest potential applicability in real world scenarios. Examples can be given as, interactive gaming, smart home systems (especially for elderly people), effective presentation possibilities, better user interfaces, (dynamic advertising, guerilla marketing) etc..

Furthermore, almost all of the metropolices around the world has a closed circuit television system (CCTV) to monitor different districts of the city. According to the British Security Industry Authority (BSIA) there are approximately 5.9 million CCTV cameras in the county of United Kingdom [6]. Even though there are automated systems that are integrated to these CCTVs, such as plate recognition and facial recognition systems with satisfactory recognition accuracies, there is not a reliable human action recognition frame-



(a) Multiple CCTVs are employed to perform surveillance  
 (b) For performing surveillance task a CCTV personel has to check multiple screens for hours in order to recognize any suspicious behaviour

Figure 1.4: Cameras and a CCTV station

work that is trusted as much as the mentioned ones.

There are also applications in the medical field [7]. During rehabilitations and physical therapies subjects behaviours are analyzed and assisted to increase the efficiency of the movement. The work of Venkataraman et al. [8] is resulted with a home rehabilitation system for patients who survived with a stroke. They claimed that every year 15 million people suffer from a stroke. Their system tracks the body movements of the patients and guide them by performing repetitive tasks for the therapy. Sung et al. [9] proposed an indoor surveillance system for elderly people in order to check their daily activities.

In some of the sports activities, trainers are used to track their athletes performances using human behaviour analysis techniques. Li et al. [10] proposed a work for automatically investigating complex diving action in challenging dynamic backgrounds. They obtained joint angles of the athlete and performed an analysis or a comparison of the athletes overall performance.

An offline application, video categorization also makes use of human ac-



tion recognition approaches. Due to the vast amount of data collected by popular video sharing websites, e.g. YouTube, Vimeo, Dailymotion etc., it becomes essential to categorize videos in terms of their content [11, 12]. In this context, Ullah et al. [13] proposed a supervised approach to learn local motion features “actlets” from annotated video data. They characterized actions with respect to joint features and their motion patterns.

## 1.2 Thesis Contributions and Organization

The main contribution of this thesis is to propose a new feature extraction and representation technique for human action recognition using depth images. We make an analogy between depth and intensity images and calculate 2D optical flows from the depth image sequences in order to capture 3D local motion variations throughout an action. Before binning the Histogram of Oriented Optical Flows from Depth (HOOFD), for data reduction purposes we define local patches around each joint by using tracked 3D skeleton data and extract optical flows from those patches. Although this step generates features that contain 3D local motion variations around each joint of an entire action sequence, it does not have sufficient temporal content. To capture the temporal evolution of the optical flow vectors, we partition the action instance into a pyramidal structure and bin the HOOFD at each level separately. Thus, temporal information of the 3D local motion vectors are injected into the feature descriptor.

The organization of this thesis is as follows: In Chapter 2 related works regarding human action recognition are presented. These works are divided into three categories, which are intensity based methods, depth map based methods and skeletal data based methods respectively. Chapter 3 details

our approach to human action recognition using depth images. In particular, the notion of depth data is described, and feature extraction and representation along with classification methods used in the thesis are presented. Experimental results and discussions are provided in Chapter 4. Chapter 5 concludes the thesis with several remarks and indicates possible future research directions.

# Chapter II

## 2 Related Work

In the literature, there are several techniques proposed for human action recognition. Most promising ones are collected and compared in the latest surveys [14–17].

Earlier works were focused on recognizing human actions from video sequences captured by RGB cameras and some of them employed spatio-temporal interest points (Cuboids [18], STIP [19]). These were statistical methods which rely on sparse features used to represent actions and they are view-invariant and robust to noise due to the characteristics of those features (Fig 2.1).

### 2.1 Intensity based methods

Yilmaz et al. [20] proposed a method that model both shape and motion of the subject. Sequence of 2D contours were extracted and formed spatiotemporal volumes (STV). To classify actions they analyzed these STVs with respect to differential geometric surface properties. A similar method was proposed by Gorelick et al. [21]. Their method was faster and did not require video alignment. In order to predict actions they employed a descriptor based on a solution to a Poisson equation. On the other hand, silhouettes that are generated by extracting foreground regions of a person image were

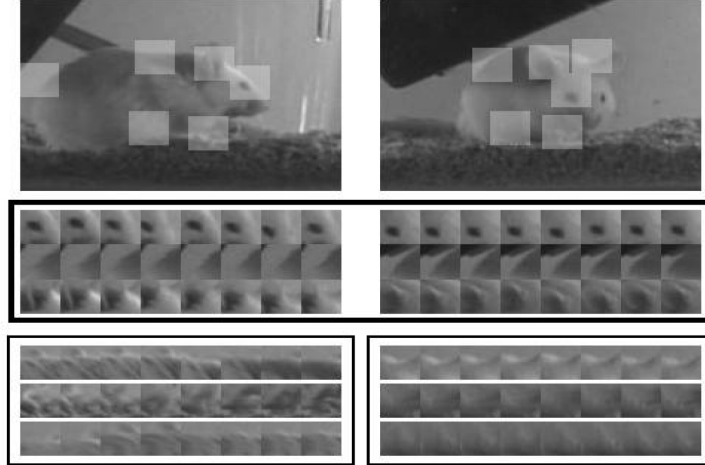


Figure 2.1: Extraction of cuboids from two different action instances, even though the posture of the mouse is quite different, extracted cuboid patches are similar [18]

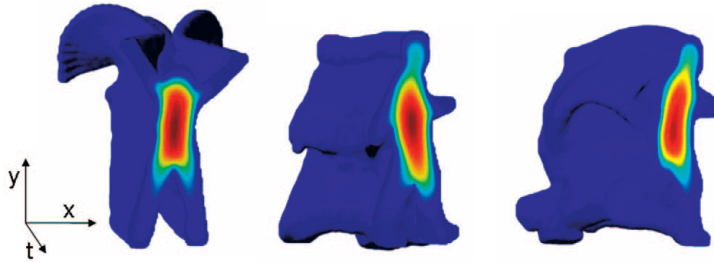
stacked consecutively to analyze surface changes of the spatio-temporal volume [22]. These sequences form Motion History Images (MHI) and Motion Energy Images (MEI) [23], which were employed as feature descriptors for template matching (see Figure 2.2).

As a grid based approach, Ikizler and Duygulu [24] used oriented rectangular patches in order to bin a grid. Each cell of the grid contained a histogram that shows the orientation distribution of the rectangular patches. Nowozin et al. [25] had a different approach. Rather than using a spatial grid as in [24], they used a temporal grid for feature representation. After extracting vectors around each interest points, they applied Principal Component Analysis (PCA) and clustered them using K-means algorithm in order to construct a codebook.

In the work of Mikolajczyk et al. [26] shape and motion features were extracted in each frame. Then, they employed the center of mass of the



(a) Motion History Images



(b) Motion Energy Images

Figure 2.2: MHI and MEI notions proposed in [21]

subject, who performed the action instance. In feature representation step they clustered these features and represented them as vocabulary trees.

Another approach is employed by Song et al. [27], they tracked points in each frame and fitted them to a triangulated graph to do the classification.

There are also remarkable works on human pose estimation from 2D still images [28]. Although these algorithms produce successful pose estimation results, they can not be used in an action recognition framework due to their significant processing time (approximately 6.6 sec).

With the release of the low-cost RGBD cameras, it has become easier to capture depth image sequences in real-time. Thus, there is an increasing research interest towards human action recognition using depth-data. Methods can be divided into two classes as in [29] which are depth map based

methods and skeletal data based methods respectively.

## 2.2 Depth map based methods

Most of the depth based methods employs spatio-temporal features. Depth data provides better understanding of the scene and the motion in the field of view. It is also invariant to sudden lighting changes. Li et al. [30] used an action graph to model the dynamics of human actions. They made use of bag of 3D points approach to characterize salient postures. These postures correspond to the nodes in the action graph. They aimed to characterize 3D shape of the salient postures with a small number of 3D points. Then, in order to capture the distribution of the points, a Gaussian Mixture Model (GMM) was fitted. Additionally authors collected a dataset, which is later called MSRAction3D and achieved 74.6 % overall recognition accuracy. The disadvantage of this method was the lack of correlation between the extracted interest points.

Yang et al. [31] proposed the feature Depth Motion Maps (DMM) to capture human actions. A DMM is generated by projecting depth frames onto three pre-defined orthogonal planes. Histogram of Gradients (HOG) were extracted from resulting depth motion maps and concatenated to generate final feature vectors. Zhang et al. [32] proposed another local spatio-temporal descriptor, which is generated by extracting intensity and depth gradients around selected feature points. For dimension reduction purposes k-means clustering was applied to the collected data and as a result a codebook was generated. To be able to perform prediction, Latent Dirichlet Allocation model (LDA) is used.

Oreifej et al. [33] proposed a new action descriptor, histogram of oriented

4D normals (HON4D) to represent depth sequence as a function of space and time. To make the descriptors more discriminative, they quantized the 4D space using the vertices of a polychoron (see Figure 2.3).

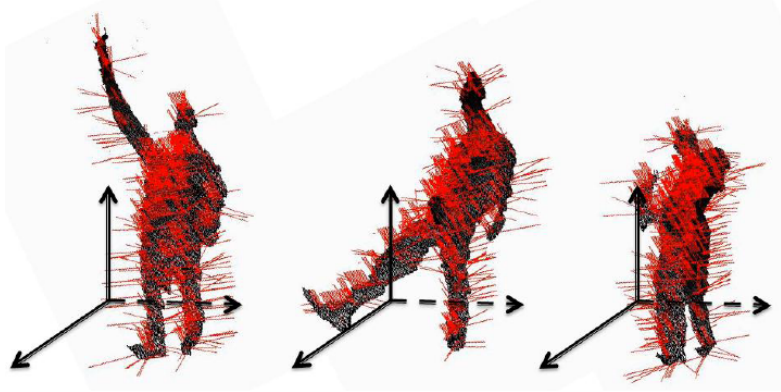


Figure 2.3: Generated 3D surface normals are illustrated in the work of [33]

In the recent work of Xia et al. [34], STIPs (Spatio-Temporal Interest Points) are extracted from depth image sequences (DSTIP). They propose a depth cuboid similarity feature (DCSF) to describe the 3D local variations around each DSTIP.

Shotton et al. [35] proposed a method to estimate 3D joint locations of a person from a single depth image. This eases the emergence of such methods since it provides real-time skeletal-data of a person.

### 2.3 Skeletal data based methods

An example of these types of works was proposed by Yang et al. [36]. They presented a new feature descriptor “EigenJoints” which combines 3 different action information; static posture, motion property and overall dynamics. In order to eliminate noise and perform data reduction, they used Accumulated Motion Energy (AME) method. Ohn-Bar et al. [37] characterize actions

using pairwise similarities between joint angle features over time. They also proposed a new feature descriptor called  $HOG^2$  which is derived by applying HOG in spatial and temporal dimensions, respectively. Lv et al. [38] designed features based on single or multiple joints. They claimed that by splitting the body parts into 3 subsets, namely leg and torso, arm, and head, they increased the discriminative power of the feature vector. Then, HMMs are built for each feature and action to be able to preserve temporal information. They also employed a multiclass Adaboost [39] classifier by combining each weak HMM classifier.

Another remarkable action recognition representation is proposed by Offi et al. [40] which is called Sequence of the Most Informative Joints (SMIJ). At each time step they compared the joints in terms of their informativeness. A joint is the most informative one, if it has the largest variance or mean among entire action instance (see Figure in 2.4). They sorted these joints with respect to their information content and generated corresponding feature vectors. 1-nearest neighbor (1NN) and SVM methods were then employed to perform the recognition step.

Vemulapalli et al. [41] represented sequence of human postures as points in the Lie Group  $SE(3) \times \dots \times SE(3)$ . This feature description modeled the 3D relationship between body parts using rotations and translations. After modeling all action instances as curves, they applied temporal modelling and classified actions using the Lie Algebra.

A different approach was proposed by Lillo et al. [42]. They modeled activities in a hierarchical manner. At the bottom level they encoded body postures using skeletal data provided by [35] and formed a dictionary of body poses. At the intermediate level in order to describe action primitives (atomic



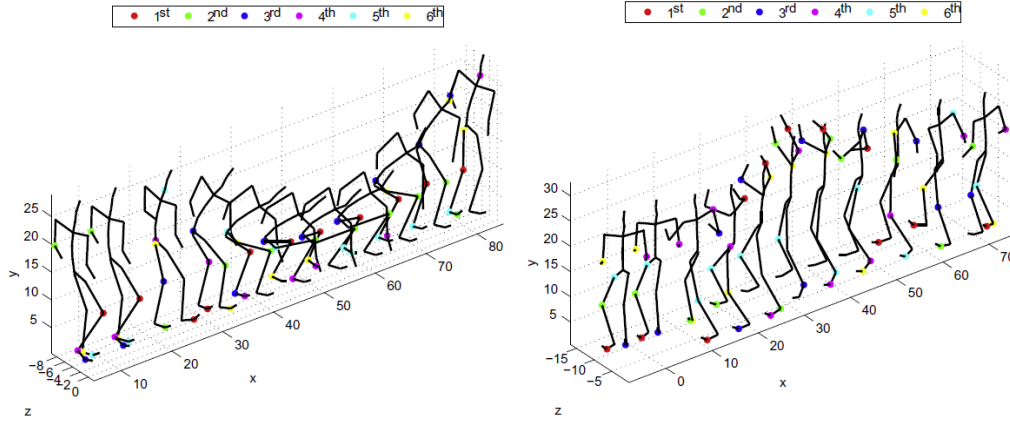


Figure 2.4: Illustration of the most informative joints during an action instance [40]

actions) they used bag-of-words (BoW) representation. Basically, they bin a histogram to model an action instance which consists of multiple sequentially body posture words. At the third level they combined these action primitives and composed complex activities.

Recently, Gupta et al. [43] propose a new approach to cross-view action recognition using 3D Mo-Cap data. Using unlabeled skeletal data, 3D posture sequences (3D joint trajectories) are recovered. Then they match these posture sequences without any need of an annotated data. Additionally, they also proposed a motion-based descriptor that is capable to compare 3D motion capture data with a 2D video data directly.

In recent works there is a trend towards the fusion of spatiotemporal and skeleton information. It implies generation of highly discriminative features.

Xia et al. [44] proposed Histogram of 3D Joint Locations as a feature. They mapped cartesian joint location coordinates into spherical coordinates  $(r, \theta, \phi)$  to satisfy view invariance. After performing linear discriminant analysis (LDA) on the feature vectors to reduce the dimension, they clustered

every posture into  $k$  visual words. Finally, in the recognition step they used discrete HMM for classification.

Zhu et al. [45] also followed this trend in their work. As a spatio-temporal feature they combined several methods and selected the ones that performed the best. These are Harris3D detector [46], Hessian Detector, HOG/HOF descriptor [47], HOG3D descriptor [48] and lastly ESURF descriptor [49]. By using skeletal data, they extracted three different features, namely pair-wise joint distances in each frame, joint location differences between subsequent frames and joint location differences between current frame and the first frame. They then applied k-means clustering to perform feature quantization. In order to perform the fusion at feature level, they proposed to use Random Forest method.

Chaaroui et al. [50] combined skeletal data and silhouette based features to utilize human action recognition. A method similar to pair-wise joint differences is employed with a different normalization scheme. They proposed a radial scheme as in [51] to obtain silhouette based features. Bag of poses method is used to perform discriminative and low dimensional feature representation. In the recognition step they used dynamic time warping (DTW) to be able to find similar action instances.

Another interesting method is proposed by Wang et. al. [52]. They calculated a combination of appearance features for each frame, namely local occupancy patterns (LOP) and pairwise relative position features of each joint. To represent temporal variation, they recursively divided the action instance into parts and generated a pyramid where short Fourier transforms were applied for all levels of the pyramid separately. Their results show that they generated sufficiently enough discriminative features to perform

the classification.

Luo et al. [53] proposed a framework that employs sparse coding and temporal pyramid matching methods for recognizing human actions. For classification stage they presented a class-specific dictionary learning algorithm. It holds the best recognition result in MSRAction3D dataset. This work is a good example of how feature representation and classification techniques affect the performance of the overall system.

A similar method with HON4D [33] was proposed by Yang et al. [54]. They collected low-level polynormals in each spatio-temporal grid. These polynormals are local clusters of extended surface normals. In addition to that, they also proposed an adaptive spatio-temporal pyramid in order to capture spatial and temporal information precisely. To represent these features, they used sparse coding and learned a dictionary accordingly.

Recently, Lu et al. [55] proposed a new feature which is called “range-sample”. This is a binary descriptor, which is generated by using the  $\tau$  test in [56]. Final binary descriptor is formed by concatenating a set of  $\tau$  test results of randomly sampled pixel pairs on a patch. It is claimed that this binary descriptor or similar ones such as HOG and SIFT are powerful due to their characterization of local edge structures. The steps of this method are as follows. First, they perform an estimation of human-activity depth range. Then, they partition the depth map into three layers, namely background, activity and occlusion layer. At the classification step they first cluster feature vectors and use them for training SVM classifiers.

Another recent approach is proposed by Lin et al. [57] for recognizing actions in RGB videos. By collecting 3D depth and skeletal data using a Kinect, they formed a database. By employing this dataset, they enhance

the capabilities of their descriptor and used it to classify actions from 2D intensity images.

# Chapter III

## 3 Action Recognition using Depth Data

Proposed technique is explained in four subsections, which is also represented as a flow chart in Figure 3.1. First, depth data is acquired using Kinect sensor and some operations are performed on the acquired data to be used by our algorithm. Then, two different features are extracted, 3D joint features and HOOFD (Histogram of Oriented Optical Flows from Depth) features. These features are represented using two different techniques: signal warping and temporal pyramid. In the classification step, Naive Bayes and Support Vector Machines classifiers are used to recognize human actions.

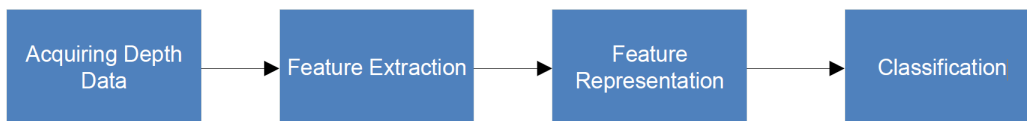


Figure 3.1: Flow Chart of the Proposed Method

### 3.1 Acquiring Depth Data

Depth sensor Kinect provides 640x480 RGB and Depth images in 30 frames per second. Its working range of depth is from 0.8 to 4 meters with an angular field of 57 degrees horizontally and 43 degrees vertically. The depth acquisition method is named as “Light Coding” which the company Prime-

sense has patented [58, 59]. Objects that are too near or too far away are shown as a black pixel on depth images (raw depth value of 2048).

There are some issues that should be taken into account before using depth data for any application. These are:

- Formation of shadows in depth data
- Eliminating the noise

The occurrence of shadows in depth data are caused by the depth measurement system of the sensor. The measurement is done with a triangulation method. IR transmitter constantly emits rays to the scene and these rays are reflected when they encounter with an object. Then, IR camera capture these reflected rays. By calculating the roundtrip time of a ray, the distance between the camera and the object is provided (see Figure 3.2). When the IR rays are obstructed by an object, IR camera cannot capture any ray from the corresponding region on the background. Thus, shadows are formed as a reflection of the object on the background.

On the other hand rough object boundaries caused noise on depth data. Therefore some regions are inaccurate, contains gaps and holes. In order to eliminate this noise, bilateral filters are used. The idea of bilateral filter is first proposed by Tomasi et al. [60]. It is a nonlinear filter employed both in spatial and range domain. It can also be interpreted as a Gaussian filter that has no effect across edges (sudden lighting changes).

$$F_{bilateral} = \frac{1}{c_n} \sum s(\|p - q\|)r(\|I_p - I_q\|) \quad (1)$$

First term in equation (1) “ $s(\|p - q\|)$ ” is referred as the space term and the second term “ $r(\|I_p - I_q\|)$ ” is referred as the range term.  $c_n$  is a normal-

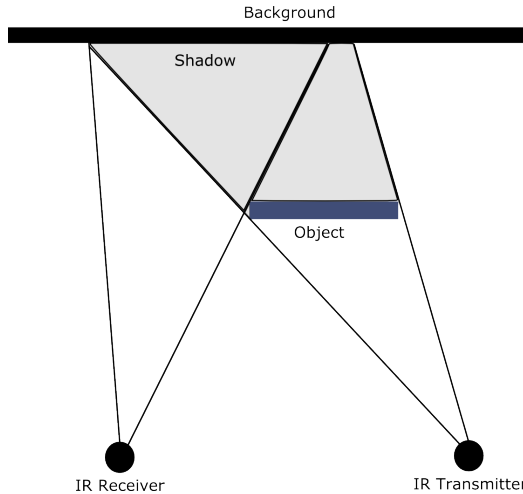


Figure 3.2: Illustrating the cause of shadow

ization factor and  $I_p$  is the intensity value of pixel  $p$  in the input image  $I$ . After applying this filter, depth data can be used for feature extraction.

### 3.2 Feature Extraction

While searching for a robust and rich feature descriptor for an action recognition framework, we observed that 3D joint locations or joint angles were not discriminative enough to represent an entire action. Even though spatial relations were encoded to the descriptor by using these features, e.g. pairwise affinities [37] or LOP [52], they do not carry the temporal information. Furthermore, a single action can be executed quite differently (in space and time). Thus, intra-class variations arise; for example one person can bend towards the camera and other can bend away from the camera. Besides one person can complete bending action in 10 seconds and another one can be faster and finish it in 5 seconds. Different solutions such as Dynamic Temporal Warping (DTW) [61] and Fourier Temporal Pyramid [52] have been

proposed to handle such cases.

In this thesis two different feature sets are used to classify human actions from depth data. First, joint features, e.g. joint angles, joint angular velocities, joint positions, and their linear velocities are calculated and investigated in terms of their performance. Next, as a new feature extraction and representation method, Histogram of Oriented Optical Flows from Depth (HOOFD) is proposed.

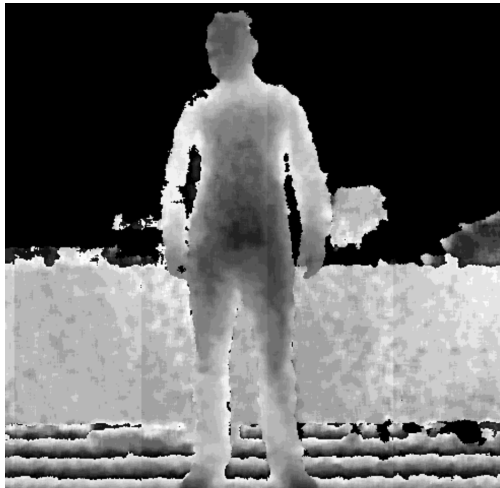
### 3.2.1 Joint Features

As mentioned before, Shotton et al. [35] proposed a human pose estimation algorithm using depth data and achieved satisfactory results. It provides real-time skeleton data of the subjects. Skeleton data consist of joints that belong to L/R foot, L/R ankle, L/R knee, L/R hand, L/R wrist, L/R elbow, L/R shoulder, neck, head, hip center, spine and shoulder center, respectively.

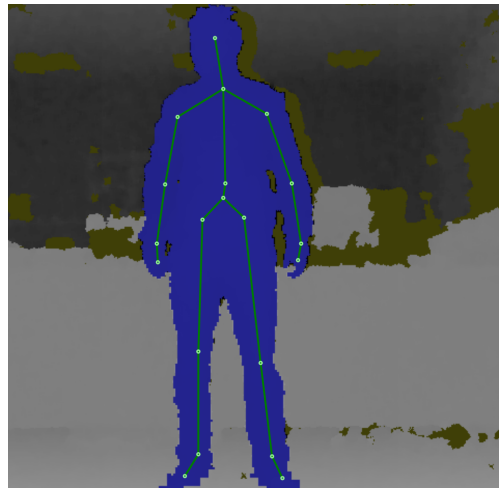
While extracting features for an action recognition framework, it is important to choose features that ensure scale and view invariance. Scale-invariance brings the advantage that even if different persons with various physical conditions (thin/overweight,tall/short etc.) performed the actions, it would not affect the systems recognition performance. To increase robustness, these invariances should be guaranteed.

Inspired from [44], 3D joint coordinates  $P_{posture} = \{p_1, \dots, p_{20}\}$  where  $p_n = (x_n, y_n, z_n)$  they are mapped to spherical coordinates  $s_n = (r, \theta, \phi)$  for a better and compact representation. We exclude the radius parameter  $r$  to gain scale-invariance. In addition to that, to remove the effect of view variance between the action instances, the origin of this spherical coordinate system is aligned with the person’s hip center as illustrated in Figure 3(c).

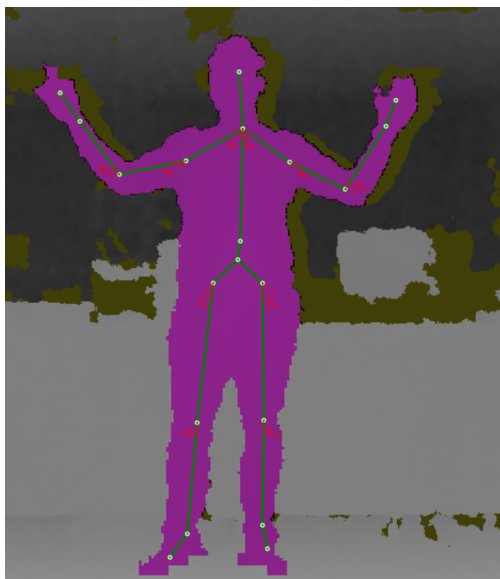




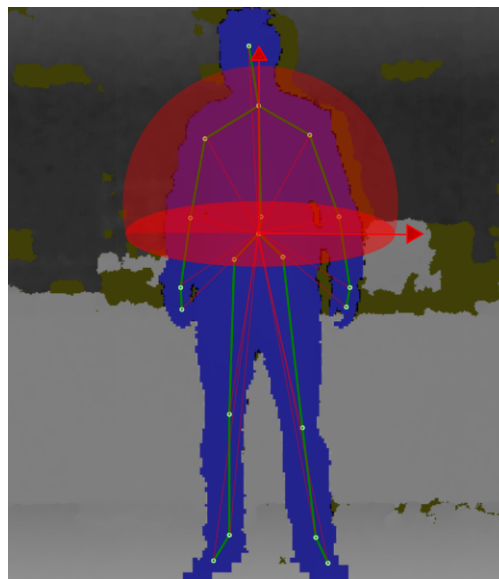
(a) Depth Data of a Person



(b) Extracted Skeleton Data



(c) Calculated Joint Angles



(d) Reference vectors and spherical coordinate system

Figure 3.3: Joint Features Illustration

Thus, instead of representing a posture with  $3 \times 20 = 60$  parameters we reduce it to  $2 \times 19 = 38$  while providing scale and view-invariance.

Furthermore, after extracting joint locations we calculate 10 joints angles

from these 20 joints. Let  $u = (u_x, u_y, u_z)$  and  $v = (v_x, v_y, v_z)$  be vectors in 3 dimensional space defining a skeleton line segment between 2 joints two joints (see Figure 3.4). For example,  $u$  can be the vector that connects shoulder and elbow joints and  $v$  can be the vector that connects elbow and hand joints. The angle between these two vectors is calculated as follows:

$$\Theta_{uv} = \tan^{-1} \left( \frac{\|u \times v\|}{u \cdot v} \right) \quad (2)$$

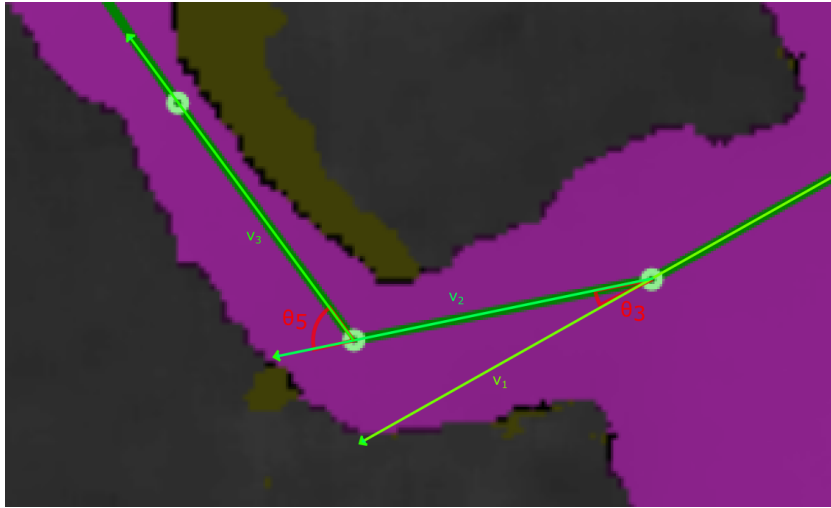


Figure 3.4: An illustration of joint angle calculation by defining vectors between each joint locations

Once joint angles are determined, their time derivatives can be computed in an approximate fashion using consecutive frames. Joints are then sorted based on their approximate velocities. While constructing the feature vector, both joint angles and joint velocities are concatenated. Their performance will be compared and investigated in Chapter 4.

### 3.2.2 Optical Flow from Depth Data

Optical flow is a motion estimation technique to calculate each pixel's independent motion using 2D intensity images. Common assumption of this estimation is that pixel intensities are translated from one pixel to the next continuously (brightness constancy constraint). As a result, an approximation of 2D motion field (projection of 3D motion field) is achieved. Brightness constancy constraint (BCC) is formulated by the following condition:

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \quad (3)$$

A depth image contains 3D world coordinates  $(x, y, z)$  of the scene points with respect to the camera frame. We make the important observation that the depth values ( $z$ ) can be represented as an intensity image. Thus, a grayscale image can be produced from a depth image by mapping depth values ( $z$ ) to 8-bit integers  $[0, 255]$  (Figure 3.5).

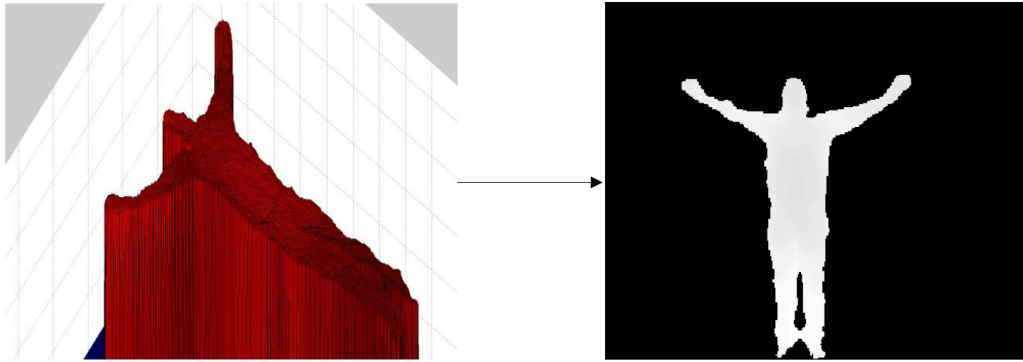


Figure 3.5: Mapping depth data to grayscale intensity image

Since we have produced a grayscale image we can now perform a 2D optical flow analysis on the resulting images. In this work, Horn-Schunck's global method [62] is employed to compute optical flow components  $(u_x, u_y)$ .

However, it should be noted that other optical flow techniques such as Lukas-Kanade (LK) [63] can be used for the same purpose.

By embedding depth information as a pixel intensity we strengthened the output of the optical flow calculation in a classification perspective. As an output, we are able to generate a feature, which is invariant to sudden change of brightness.

### 3.3 Feature Representation

In most of the works, recognition accuracies are strongly dependent on its feature representation technique. In this thesis we used two different approaches to represent our features. These are signal warping, and patch extraction and temporal pyramid.

#### 3.3.1 Signal Warping

This method is used for the joint features in section 3.2.1. Due to the varying time intervals of action instances, signal warping is chosen. For each experiment a global action instance interval is set. An action instance  $S$  with  $n$  number of frames is basically warped and its duration (number of frames) is increased/decreased with respect to the assigned global variable. It is done by randomly replicating some of the frames in the frame sequence and concatenating one another (see Figure 3.6).

#### 3.3.2 Patch Extraction and HOOFD Features

The joint location estimation algorithm [35] provides 20 2D/3D joint location coordinates from depth data in a very precise manner. In our work, we use only 10 joints, namely L/R shoulder, L/R elbow, L/R knee, L/R wrist,

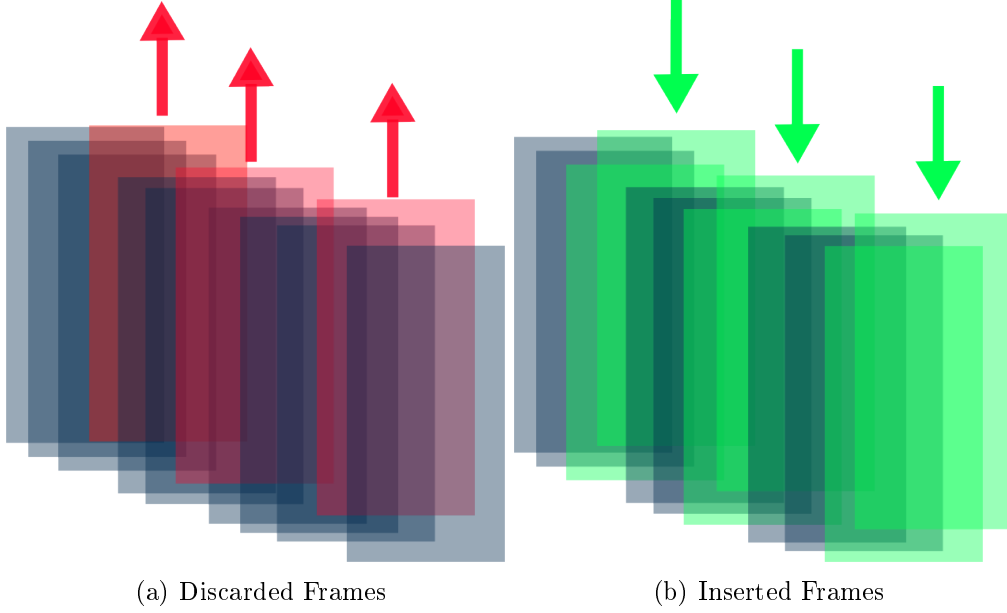


Figure 3.6: Randomly selected frames are discarded / replicated and inserted in an action sequence

shoulder center and head, to extract  $m \times m$  (typical value of  $m$  is 11) patches around each of them.

During our experiments we observed that remaining joint locations were less robust to noise and viewing conditions than the selected ones.

Optical flow patches of joint  $J$  in frame  $i$  are defined as:

$$P_{J,u_x}^{(i)} = \begin{pmatrix} u_{x,1} & \cdots & u_{x,m} \\ u_{x,m+1} & \cdots & u_{x,2m} \\ \vdots & \ddots & \vdots \\ u_{x,m(m-1)} & \cdots & u_{x,m^2} \end{pmatrix}, \quad P_{J,u_y}^{(i)} = \begin{pmatrix} u_{y,1} & \cdots & u_{y,m} \\ u_{y,m+1} & \cdots & u_{y,2m} \\ \vdots & \ddots & \vdots \\ u_{y,m(m-1)} & \cdots & u_{y,m^2} \end{pmatrix} \quad (4)$$

After calculating optical flows  $(u_x, u_y)$  from depth images and extracting patches  $P_J$  around each joint  $J \in \{1, \dots, 10\}$ , these optical flow patches are

concatenated for the entire action sequence. Next, a novel depth feature, Histogram of Oriented Optical Flows from Depth (HOOFD) is proposed. In order to calculate HOOFDs using concatenated optical flow patches, a similar procedure to [64] is used.

An orientation image  $\theta$  and a magnitude image  $M$  are calculated by

$$\theta = \text{atan2}\left(\frac{u_y}{u_x}\right) \quad , \quad M = \sqrt{u_x^2 + u_y^2} \quad (5)$$

These images are used to bin a histogram based on two features, the primary angle between the flow vector and the horizontal axis, and magnitude of the flow vector. While constructing the histogram, this combination encodes both the direction and the magnitude of the flow vectors. The contribution of each vector to its corresponding bin is proportional to its magnitude.

Inspired from temporal Fourier Pyramid reported in [36], we construct a new feature as Pyramidal Histogram of Oriented Optical Flows from Depth to capture temporal motion information.

Pseudo code of the Pyramidal HOOFD construction algorithm is given below.

Additionally pyramidal feature construction for a 2-level pyramid can be illustrated as follows.

At the first level, feature vector  $F_{L1}$  of the entire action instance is calculated:

$$F_{L1} = \text{HOOFD}(P_{J,u_x}^{(1:n)}, P_{J,u_y}^{(1:n)}) \quad (6)$$

**Algorithm 1: PYRAMIDAL HOOFD FEATURE CONSTRUCTION**

**Input:** Joint Patches  $P_J\langle P_J^{(1)}, P_J^{(2)}, \dots, P_J^{(n)} \rangle$  & number of pyramid levels  $L$

**Output:** Feature vector  $F$  that is generated by concatenating HOOFD outputs at each level respectively

```

level ← L;
for each joint  $J$  do
     $V_x = \text{concat}(P_{(J,u_x)}^{(1)}, P_{(J,u_x)}^{(2)} \dots, P_{(J,u_x)}^{(n)});$ 
     $V_y = \text{concat}(P_{(J,u_y)}^{(1)}, P_{(J,u_y)}^{(2)} \dots, P_{(J,u_y)}^{(n)});$ 
    for each level do
        Divide the vectors into  $2^{\text{level}-1}$  parts;
        Calculate HOOFD of each part
    end
    Concatenate resulting histograms into  $F$ 
end
return  $F$ ;

```

At the second level, sequences with the length of  $n/2$  are employed.

$$F_{L2,1} = \text{HOOFD}(P_{J,u_x}^{(1:n/2)}, P_{J,u_y}^{(1:n/2)}) \quad (7)$$

$$F_{L2,2} = \text{HOOFD}(P_{J,u_x}^{(n/2+1:n)}, P_{J,u_y}^{(n/2+1:n)}) \quad (8)$$

The final feature vector  $F$  is constructed by concatenating the feature vectors computed at each temporal level, i.e.  $F = (F_{L1}, F_{L2,1}, F_{L2,2})$ . General overview of the proposed framework is illustrated in Figure 3.7.

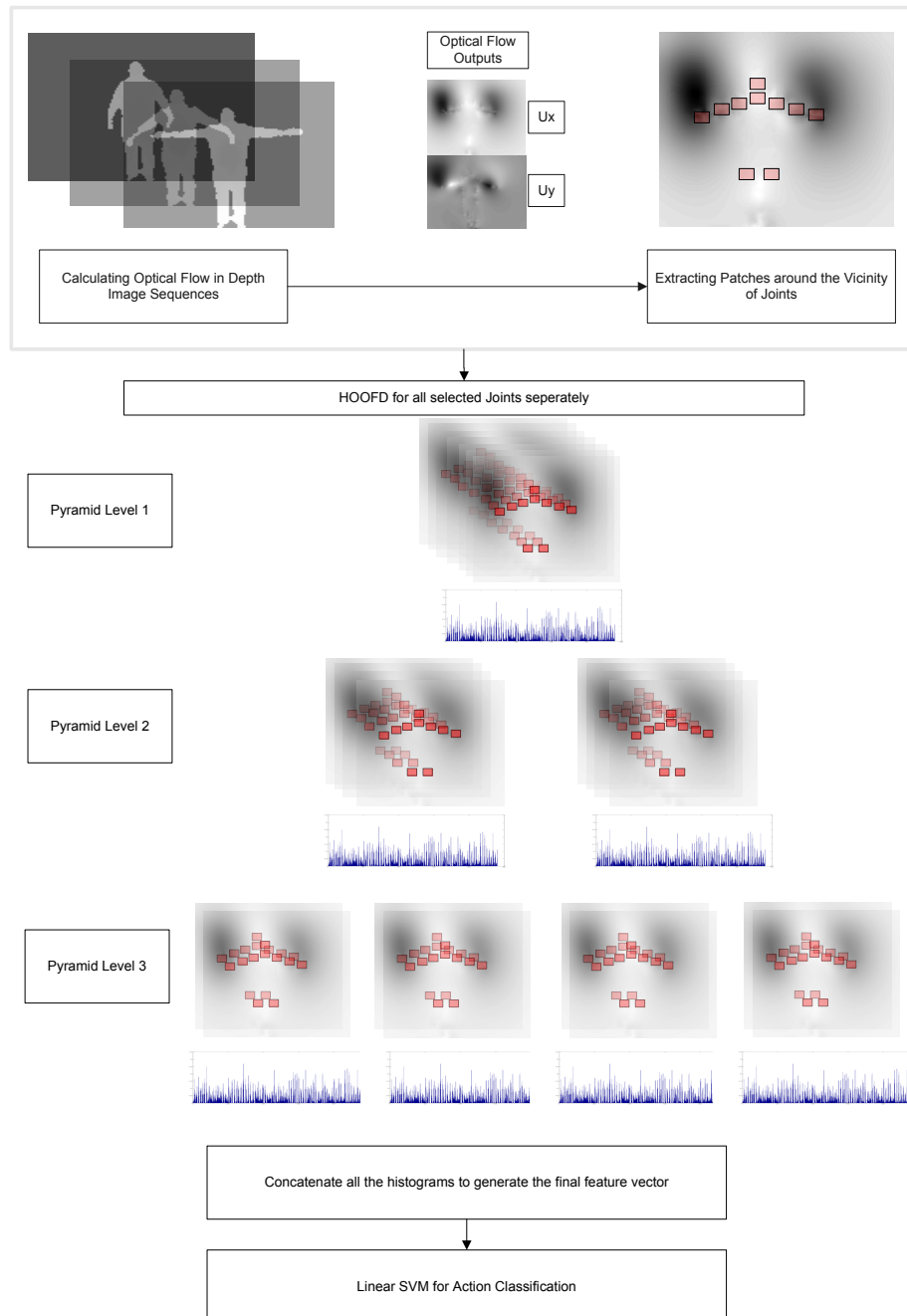


Figure 3.7: Overview of the proposed method



## 3.4 Classification Methods

### 3.4.1 Naive Bayes Classifier

Naive Bayes is a popular and supervised learning method, which works quite well in real world scenarios. Main assumption of this method is that the values of the features are independent of each other given their class (conditional independence assumption). Given data  $((x^{(1)}, y_1), (x^{(2)}, y_2), \dots, (x^{(n)}, y_n))$ , where  $x^{(i)} = (x_1^{(i)}, \dots, x_d^{(i)})$  is the feature vector and  $y_i$  is the class label. A distribution of features is interpreted as a joint probability distribution  $p(x, y) = p(x|y)p(y)$ . Due to the independence assumption conditional probability  $p(x|y)$  can be expressed as  $(p(x_1|y) \dots p(x_d|y))$ . To be able to predict a test data  $x_{test}$  we compute maximum posterior probability as follows:

$$\hat{y} = \operatorname{argmax}_{y \in Y} P(y|x) \quad (9)$$

$$\hat{y} = \operatorname{argmax}_{y \in Y} \left( \frac{P(x|y)P(y)}{P(x)} \right) \quad (10)$$

$P(x)$  does not depend on  $y$  so it is usually discarded from this calculation. Additionally during the experiments we assumed that distributions are Gaussians with identical diagonal covariance matrices.

### 3.4.2 Support Vector Machines

Support Vector Machines (SVM) is a supervised learning method defined by a separating hyperplane. Briefly, it calculates as an output the choice of the most optimal hyperplane in Figure 3.8, which is the one that possess the maximum margin from the training data. In order to find detailed information, More details can be found in books on pattern recognition such

as [65–67].

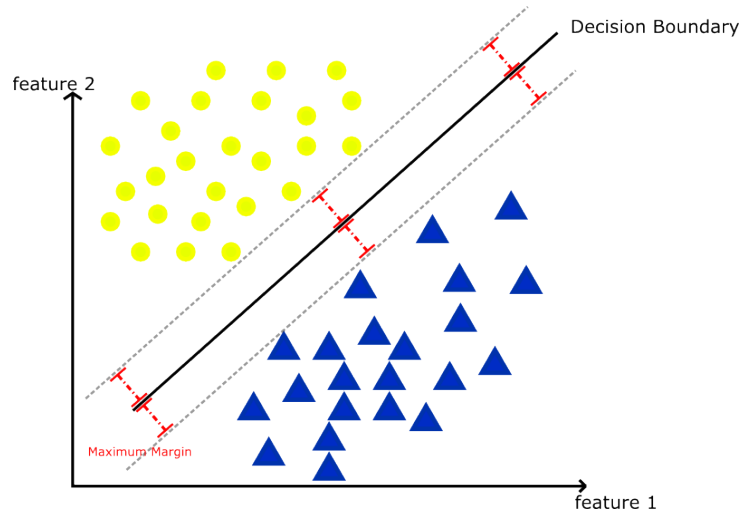


Figure 3.8: SVM classifier returns maximum margin decision boundary (hyperplane)

There are several SVMs that employ with different kernel functions, e.g. linear, polynomial, sigmoid etc.. We employed the one with a linear kernel. A popular SVM package libSVM [68] is used during our implementations.

# Chapter IV

## 4 Experiments

### 4.1 Datasets

We assessed the performance of our proposed method by conducting several experiments with publicly available human action recognition datasets, MSRAction3D [30], MSR Action Pairs 3D [33] and MSRC-12 Gesture Dataset [69].

#### 4.1.1 MSR Action3D Dataset

MSR-Action3D is a widely used action recognition dataset, which consists of depth sequences captured by Microsoft Kinect at 15 Hz, and image and world joint coordinates of each subject. An example sequence of depth images from this dataset is depicted in Figure 4.1. Dataset contains 20 actions performed by 10 subjects.

Additionally, depth images were preprocessed in order to clear background noises caused by the depth sensor. This is a challenging dataset because it includes highly similar actions. We followed the same experimental settings as in [30] and splitted the dataset into 3 subsets as shown in Table 4.1.

AS1 and AS2 represent actions with similar movements, e.g. in AS1

Action Sets		
Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Hammer (2)	Wave (1)	Throw (6)
Smash (3)	Catch (4)	Forward Kick (14)
Forward Punch (5)	Draw X (7)	Side Kick (15)
Throw (6)	Draw Tick (8)	Jogging (16)
Clapping Hands (10)	Draw Circle (9)	Tennis Swing (17)
Bend (13)	2 Hand Wave (11)	Tennis Serve (18)
Tennis Serve (18)	Side Boxing (12)	Golf Swing (19)
Pickup and Throw (20)	Forward Kick (14)	Pickup and Throw (20)

Table 4.1: Actions of MSRAction3D are divided into 3 subsets (numbers in paranthesis represents the action annotations)

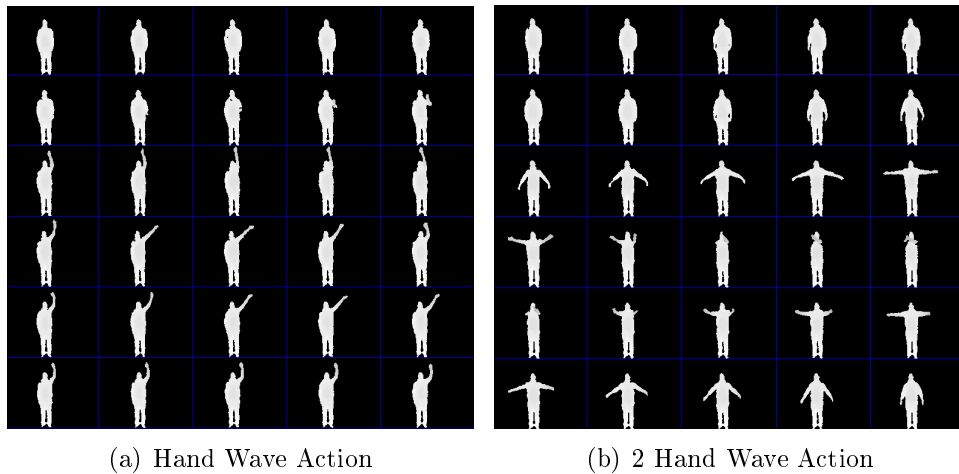


Figure 4.1: Depth image sequence examples from MSRAction3D dataset

“forward punch” and “hammer” are likely to be confused by each other. AS3 contains more complex and distinct actions.

#### 4.1.2 MSR Action Pairs Dataset

This dataset is used by [33]. There are two important differences that distinguishes this dataset from previous MSRAction3D. First, in the MSRAction

tion3D dataset, many actions are performed while subjects are standing still, thus skeletal data seems reliable to represent an entire action. Second, some actions have very similar and limited body part motions (e.g. hammer and forward punch), which reduce the reliability of extracted motion cues. For these reasons, six pairs of activities are selected in this new dataset: “Pick up a box/Put down a box”, “Lift a box/Place a box”, “Push a chair/Pull a chair”, “Wear a hat/Take off a hat”, “Put on a backpack/Take off a backpack” and “Stick a poster/Remove a poster” (see Figure 4.3). We used the same experimental settings and employed the comparison presented in [33].

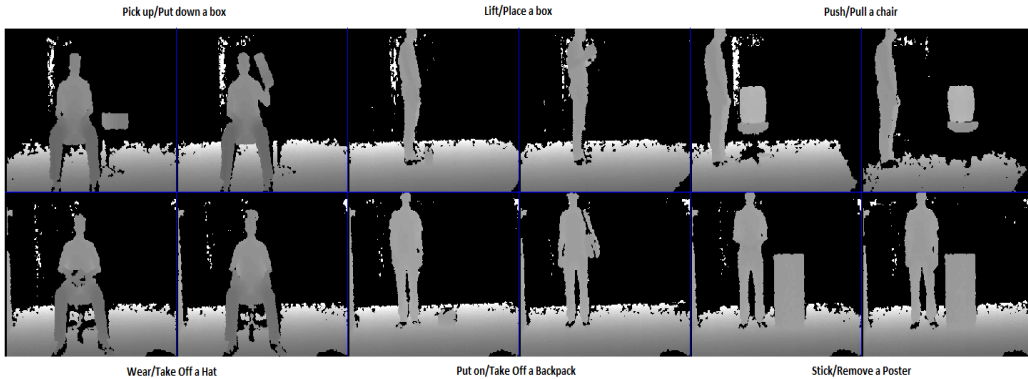


Figure 4.2: Depth image examples of MSR Action Pairs dataset

### 4.1.3 MSRC-12 Gesture Dataset

This dataset is collected at MSR Cambridge as a part of the work in [69]. It consist of 594 sequences and 719359 frames of 12 gestures (actions) that is performed by 30 subjects. However they did not share publicly the depth data of the frames, they only gave access to 3D skeletal coordinates at each frame. They categorized gestures into two sub categories, namely Iconic gestures and Metaphoric gestures. Iconic gestures are crouch, war a goggles,

shoot a pistol, throw an object, change weapon and kick. Metaphoric gestures represent a more abstract object, such as starting the music or raising the volume, where the subject lifts and outstretched him/her arms. Others are moving arm to the right, wind it up, bow, had enough and beat both. It should be noted that “had enough” action can be performed differently across subjects.

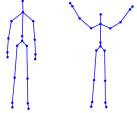
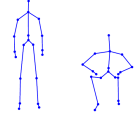
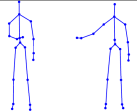
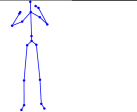
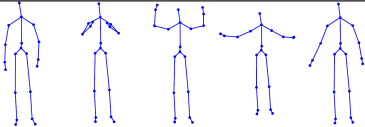
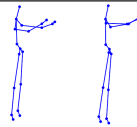
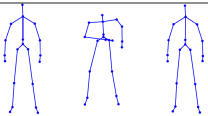
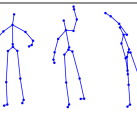

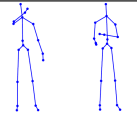
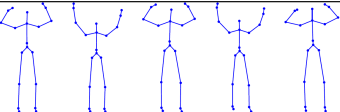
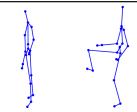
Metaphoric gestures	Main frames	Iconic gestures	Main frames
Start music\ raise volume (G1)		Crouch or hide(G2)	
Navigate to next menu(G3)		Put on night vision goggles(G4)	
Wind up the music(G5)		Shoot with a pistol(G6)	
Take a bow to end the session(G7)		Throw an ob- ject such as a grenade(G8)	
Protest the music(G9)		Change weapon(G10)	
Lay down the tempo of a song(G11)		Kick to at- tack an en- emy(G12)	

Figure 4.3: Gestures and captured frames from gesture instances of MSRC-12 dataset [70]

## 4.2 Joint Features with Signal Warping

In order to evaluate joint features, we performed several experiments by combining multiple features. We employed MSRC-12 Gesture dataset and extracted all the joint features that are mentioned in feature extraction section (Joint Angles, Spherical Coordinates, list of the most active joints and joint angle velocities). We generate 4 different feature sets as follows:

<b>Feature Sets</b>	<b>Feature Content</b>
F1	Joint Angles + Joint Locations in Sph. Coord.
F2	Joint Angles + Joint Angular Velocities
F3	F1 + Joint Angular Velocities
F4	F3 + List of the most active Joints

Table 4.2: Feature sets are generated in order to use on MSRC-12 Gesture dataset

First, employing leave-one-subject-out-cross-validation (LOSOVCV) we compare the classification accuracy results of both between features and between classifiers. This gives us an understanding about the reliability of both feature representation and selected classification methods. Then, we split the subjects into two equal subsets (1:1 in Table 4.3) . We used first 6 subjects for training the classifier and the remaining ones for prediction test. Furthermore, we split the subjects as 1:3, first 4 subjects are used for testing and the rest 8 are used for training the classifier. Results and comparisons are shown in Table 4.3.

From the results in Table 4.3 and in Figure 4.4 it can be easily concluded that feature set F2, which consists of joint angles and joint angular velocities, is the least discriminative one between the feature sets. The reason is that while calculating joint angular velocities, information is lost due to differentiating joint angles. We tested this approach because it is the most

Feature Sets	SVM			Naive Bayes		
	1:1	1:3	LOSOCV	1:1	1:3	LOSOCV
F1	76.45%	76.32%	75%	78%	76.8%	77%
F2	67.4%	67.4%	59.8%	74.5%	76.8%	72%
F3	75.16%	74.4%	72.7%	78%	77.3%	77.9%
F4	77.4%	76.8%	74.1%	77.7%	77.3%	72.8%

Table 4.3: Recognition accuracies (%) Comparison of different tests for MSRC-12 Gesture dataset

intuitive way to model human actions from the 3D skeletal joint features. Even though the recognition accuracies do not differ significantly, highest rates are achieved while employing Feature Set 1. It can be concluded that joint velocities and the list of the most active joints had almost no contribution to the output of the classification algorithm. This opinion can be further investigated by looking at the recognition rates of SMIJ feature in the work [40].

### 4.3 Pyramidal HOOFD Features

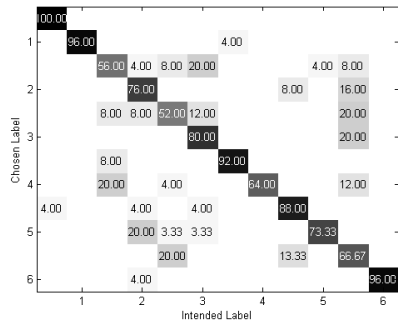
For the first experiment we carried out a Cross Subject test, half of the subjects are used for training and the rest for testing. For classification, we trained a linear SVM classifier. The results are shown and compared in Table 4.4.

CrsSubj Test	Li et al. [30]	Yang et al. [36]	Xia et al. [44]	<i>HOOFD</i>
AS1	72.9%	74.5%	87.8%	75.47%
AS2	71.9%	76.1%	85.48%	77.88%
AS3	79.2%	96.4%	63.46%	76.79%

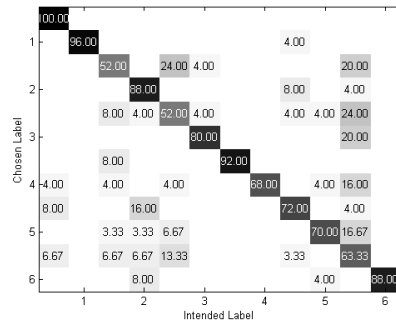
Table 4.4: Recognition accuracies (%) Comparison of Cross Subject Test for MSR Action 3D

Figure 4.5 shows confusion matrices of the classification results with re-

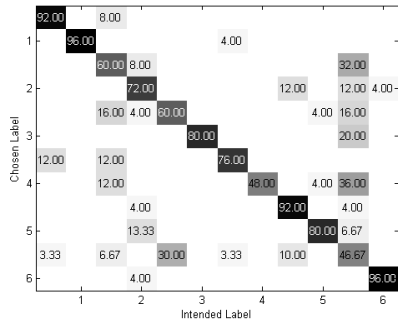




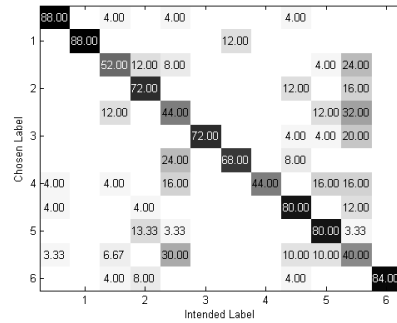
(a) F1 using Naive Bayes Classifier



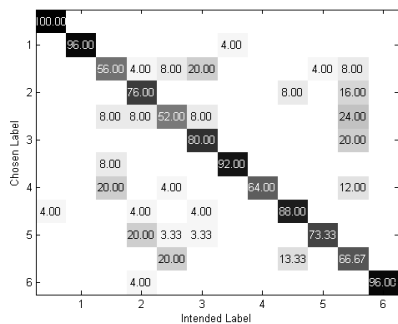
(b) F1 using SVM Classifier



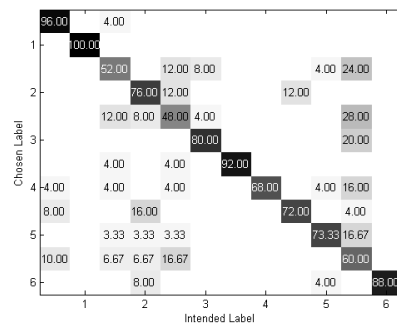
(c) F2 using Naive Bayes Classifier



(d) F2 using SVM Classifier



(e) F3 using Naive Bayes Classifier



(f) F3 using SVM Classifier

Figure 4.4: Confusion matrices of MSRC-12 dataset using (1:1) experimental settings

spect to the Action Sets. For the first Action Set AS1, it can be observed From Figure 5(a) that our classifier mismatches ‘smash’, ‘forward punch’ and ‘throw’ actions due to similar skeletal motions and local flow fields. For the second Action Set AS2, it can be seen from Figure 5(b) that we achieved 77.88% recognition rate. However, our classifier fails during categorizing very similiar actions such as ‘Catch’ and ‘Drawing’ actions. Finally, for the third Action Set AS3, Figure 5(c) shows misclassification results due to the noise on the skeletal data while subjects performing ‘Golf Swing’ action.

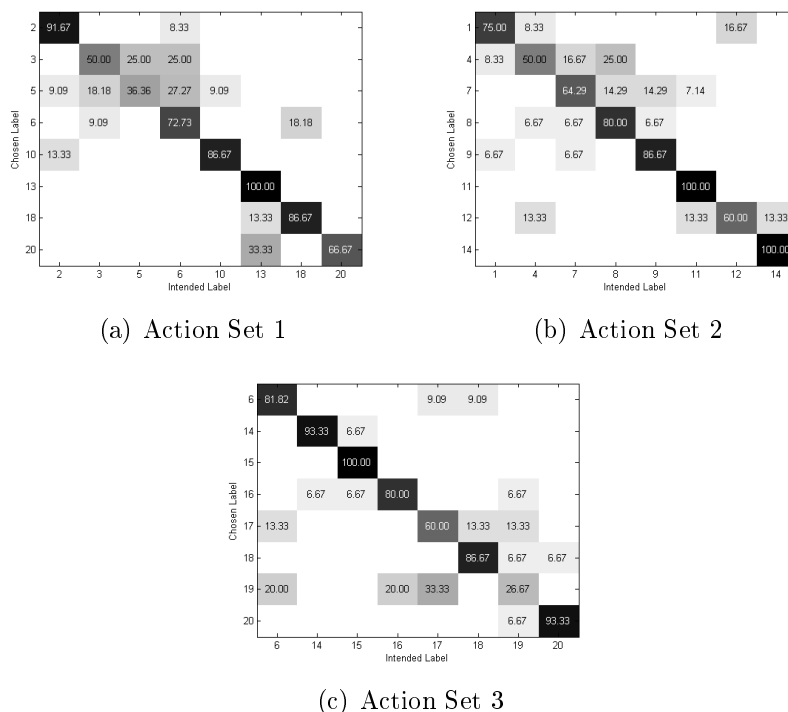


Figure 4.5: Confusion matrix of different action sets under Cross Subject Test

Table 4.5 provides another comparison of our method with the state-of-the-art algorithms. Recently proposed method [71] that uses sequence of

most informative joints achieved 47.1% recognition rate. This relatively low recognition result could be due to the noise on the skeleton tracker results as shown in Figure 4.6 and short duration of action instances. DTW [61] and HMM [38] are both typical approaches toward modeling ‘temporal’ dynamics of an action. They achieved 54% and 63% recognition rates respectively.

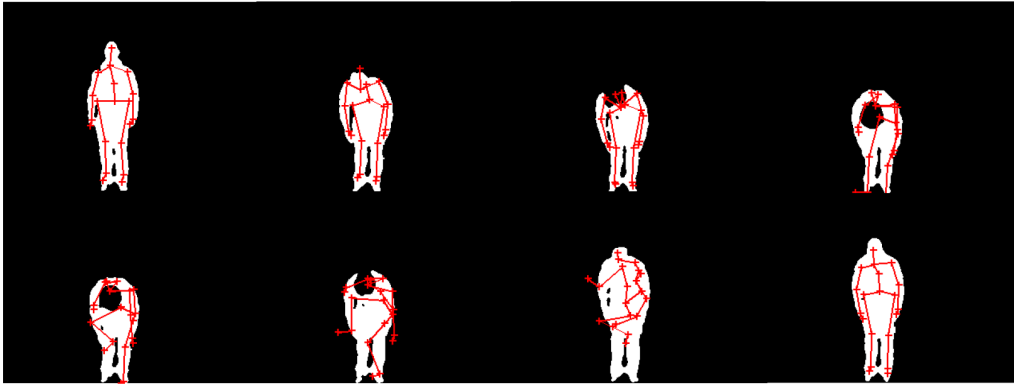


Figure 4.6: Visualization of the skeleton tracker Failure on bend action

The work in [52] used actionlet mining method and achieved 88.2% accuracy by fusing multiple features (skeleton joints and local occupancy patterns). HON4D method [33] did not employ 3D joint locations, instead they calculated the distribution of the surface normal orientations in 4D space at pixel-level. This has a high discriminative power due to its dense structure. Our method achieved 76.71% on this test, since skeleton trackers were failed drastically in some of the action instances. For that reason we could not always extract meaningful patches and this leads to relatively low classification rates.

While testing the method with MSR Action Pairs Dataset five subjects were used for training a linear SVM classifier and the rest for testing the performance of this classifier. Results are provided in Table 4.6 and the

<b>Method</b>	<b>Classification Accuracy</b>
Sequence of Most Informative Joints (SMIJ) [71]	47.1%
Dynamic Temporal Warping [61]	54%
Hidden Markov Model [38]	63%
Action Graph on Bag of 3D Joints [30]	74.7%
<b>HOOFD</b>	76.71%
EigenJoints [36]	82.3%
Actionlet Ensemble [52]	88.2%
HON4D + $D_{disc}$ [33]	88.89%
OhnBar et al. [37]	94.84%
Range-Sample Depth Feature [55]	95.62%
DL-GSGC+TPM [53]	96.7%

Table 4.5: Comparison of classification accuracy with state-of-the-art methods for MSRAction3D dataset

<b>Method</b>	<b>Classification Accuracy</b>
Skeleton + LOP [52]	63.33%
Skeleton + LOP + Pyramid [52]	82.22%
<b>HOOFD</b>	91.67%
HON4D [33]	93.33%
HON4D + $D_{disc}$ [33]	96.67%

Table 4.6: Classification accuracy comparison for MSR Action Pairs dataset

corresponding confusion matrix is shown in Figure 4.7. We compared our method with state-of-the-art methods, namely HON4D and pairwise joint affinities with local occupancy patterns (LOP). First, skeleton features and LOP were calculated in each frame of an action instance and then an SVM was trained accordingly. This method achieved 63.3% classification accuracy. Even though this feature fuses motion and shape features, it suffers from the lack of temporal relations. By employing Fourier temporal pyramid, the recognition rate was increased to 82.22%. Our method achieved 91.67% classification rate on this dataset.

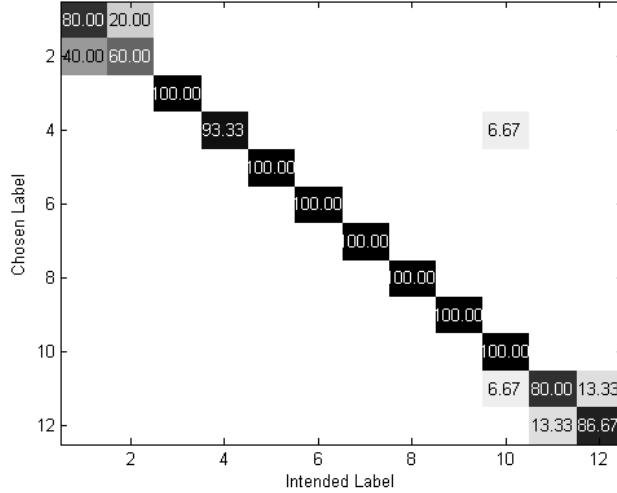


Figure 4.7: Confusion Matrix of MSR Action Pairs dataset under Cross Subject Test

Additionally, we also evaluated our algorithm with MSR Action Pairs using leave-one-subject-out-cross-validation while varying the pyramid level. Classification accuracies are tabulated in Table 4.7. These results clearly verify that employing a pyramidal approach to model the temporal variations of local motion vectors improves the classification accuracies significantly.

Pyramid Level	Feature Dimension	Classification Accuracy
Level 1	300	77.92%
Level 2	900	88.54%
Level 3	2100	93.58%
Level 4	4500	<b>95.26%</b>

Table 4.7: Classification accuracy of our method at each pyramid level

In order to investigate the effect of the patch size on the recognition accuracy, another experiment is conducted. This time action Set 2 is employed and HOOFDs are calculated with different sizes of patches (7x7, 11x11, 15x15, 21x21, 25x25, 29x29). Two different experimental settings are used

which are LOSOCV (leave one subject out cross validation) and (1:1) cross subject test. Results are shown in Figure 4.8.

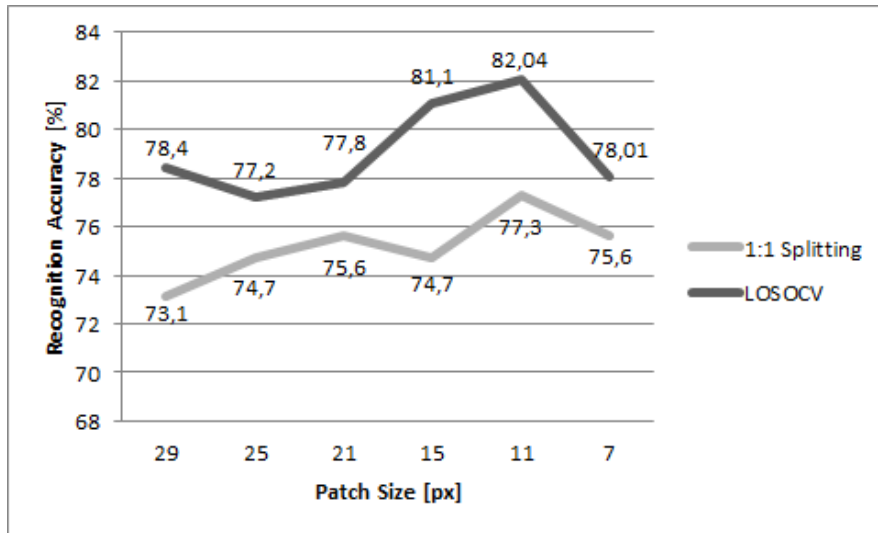


Figure 4.8: Recognition results for comparing patch size

It is clear that while increasing the area of the extracted patches around each joint, recognition accuracies tend to decrease. On the other hand, there is a lower bound of patch area which is between 9 and 13. It was found empirically during the experiments.

# Chapter V

## 5 Conclusion & Future Work

We have now presented a new human action recognition method from depth images. By drawing an analogy between depth and intensity images, we introduced a new feature called Histogram of Oriented Optical Flows from Depth (HOOFD). To reduce the dimension of the feature vectors, we used tracked 3D skeleton data to extract patches around the vicinity of each joint. After combining these patches throughout the action instance, HOOFDs are generated. To encode coarse to fine temporal variations, a pyramidal approach is utilized. Experimental results performed on three publicly available datasets and comparisons with the state-of-the-art algorithms verify the success of the proposed approach. Besides, in order to test the reliability of the HOOFDs, different experimental settings (leave-one-subject-out-cross-validation and (1:1)) and different classifiers (Naive Bayes and SVM) are employed and compared. Results prove that HOOFD features provide enough discriminativeness to represent various human actions.

Regarding future work, the potential of HOOFD features will be fully explored. Fusing HOOFD with features extracted from RGB images can enrich the resulting feature vectors. Moreover, recognition accuracies can also be increased by using pyramidal HOOFD features for training different classifiers.

## References

- [1] F. Dellaert, S.M. Seitz, C.E. Thorpe, and S. Thrun. Structure from motion without correspondence. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 557–564 vol.2, 2000.
- [2] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.
- [3] Microsoft Corporation Redmond WA. Kinect, 2013.
- [4] Leap motion controller, website, 2014.
- [5] Google. Google, project tango, 2014.
- [6] Janice Turner. Cctv britain, the world’s most paranoid nation, 2013.
- [7] Akin Avci, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. Activity recognition using inertial sensing for health-care, wellbeing and sports applications: A survey. In *Architecture of Computing Systems (ARCS), 2010 23rd International Conference on*, pages 1–10, Feb 2010.
- [8] Vinay Venkataraman, Pavan Turaga, Nicole Lehrer, Michael Baran, Thanassis Rikakis, and Steven L. Wolf. Attractor-shape for dynamical analysis of human movement: Applications in stroke rehabilitation and action recognition. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '13*, pages 514–520, Washington, DC, USA, 2013. IEEE Computer Society.



- [9] Jaeyong Sung, C. Ponce, B. Selman, and A Saxena. Unstructured human activity detection from rgb-d images. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 842–849, May 2012.
- [10] Haojie Li, Shouxun Lin, Yongdong Zhang, and Kun Tao. Automatic video-based analysis of athlete action. In *Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on*, pages 205–210, Sept 2007.
- [11] Bingbing Ni, Yang Song, and Ming Zhao. Youtubeevent: On large-scale video event classification. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1516–1523, Nov 2011.
- [12] Zheshen Wang, Ming Zhao 0003, Yang Song, Sanjiv Kumar, and Baoxin Li. Youtubecat: Learning to categorize wild web videos. In *CVPR*, pages 879–886. IEEE, 2010.
- [13] Muhammad Muneeb Ullah and Ivan Laptev. Actlets: A novel local representation for human action recognition in video. In *ICIP'12*, pages 777–780, 2012.
- [14] J.K. Aggarwal and M.S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, April 2011.
- [15] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.*, 115(2):224–241, February 2011.

- [16] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2):90–126, November 2006.
- [17] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, 2010.
- [18] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on Computer Communications and Networks, ICCCN '05*, pages 65–72, Washington, DC, USA, 2005. IEEE Computer Society.
- [19] Ivan Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, September 2005.
- [20] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 984–989 vol. 1, June 2005.
- [21] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.
- [22] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *In ICCV*, pages 1395–1402, 2005.

- [23] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):257–267, March 2001.
- [24] Nazli Ikizler and Pinar Duygulu. Human action recognition using distribution of oriented rectangular patches. In *IN: WORKSHOP ON HUMAN MOTION*, pages 271–284, 2007.
- [25] S. Nowozin, G. Bakir, and K. Tsuda. Discriminative subsequence mining for action classification. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007.
- [26] K. Mikolajczyk and H. Uemura. Action recognition with appearance-motion features and fast search trees. *Comput. Vis. Image Underst.*, 115(3):426–438, March 2011.
- [27] Yang Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(7):814–827, July 2003.
- [28] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision*, 99:190–214, 2012.
- [29] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall. A survey on human motion analysis from depth data. In Marcin Grzegorzec, Christian Theobalt, Reinhard Koch, and Andreas Kolb, editors, *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, volume 8200 of *Lecture Notes in Computer Science*, pages 149–187. Springer Berlin Heidelberg, 2013.

- [30] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010.
- [31] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM International Conference on Multimedia, MM '12*, pages 1057–1060, New York, NY, USA, 2012. ACM.
- [32] Hao Zhang and L.E. Parker. 4-dimensional local spatio-temporal features for human activity recognition. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 2044–2049, Sept 2011.
- [33] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, pages 716–723, 2013.
- [34] Lu Xia and J. K. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, pages 2834–2841, 2013.
- [35] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *Pattern Analysis and Machine Intelligence*, 35(12), 2013.
- [36] Xiaodong Yang and Yingli Tian. Effective 3d action recognition using eigenjoints. *J. Vis. Comun. Image Represent.*, 25(1):2–11, January 2014.

- [37] E. Ohn-Bar and M.M. Trivedi. Joint angles similarities and hog2 for action recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 465–470, June 2013.
- [38] Fengjun Lv and Ramakant Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV, ECCV'06*, pages 359–372, Berlin, Heidelberg, 2006. Springer-Verlag.
- [39] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting, 1997.
- [40] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, 2013.
- [41] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [42] I. Lillo, JC. Niebles, and A. Soto. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [43] Ankur Gupta, Julieta Martinez, James J. Little, and Robert J. Woodham. 3d pose from motion for cross-view action recognition via non-linear circulant temporal encoding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

- [44] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *CVPR Workshops*, pages 20–27. IEEE, 2012.
- [45] Yu Zhu, Wenbin Chen, and Guodong Guo. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image and Vision Computing*, 32(8):453 – 464, 2014.
- [46] Chris Harris and Mike Stephens. A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988.
- [47] I Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.
- [48] Alexander Kläser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 995–1004, sep 2008.
- [49] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008. Similarity Matching in Computer Vision and Multimedia.
- [50] AA Chaaoui, J.R. Padilla-Lopez, and F. Florez-Revuelta. Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 91–97, Dec 2013.

- [51] Alexandros Andre Chaaaraoui and Francisco Flórez-Revuelta. Human action recognition optimization based on evolutionary feature subset selection. In *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation, GECCO '13*, pages 1229–1236, New York, NY, USA, 2013. ACM.
- [52] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Learning actionlet ensemble for 3d human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(5):914–927, May 2014.
- [53] Jiajia Luo, Wei Wang, and Hairong Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1809–1816, Dec 2013.
- [54] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, June 2014.
- [55] Cewu Lu, Jiaya Jia, and Chi-Keung Tang. Range-sample depth feature for action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [56] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10*, pages 778–792, Berlin, Heidelberg, 2010. Springer-Verlag.
- [57] Yen-Yu Lin, Ju-Hsuan Hua, Nick C. Tang, Min-Hung Chen, and Hong-Yuan Mark Liao. Depth and skeleton associated action recognition with-

- out online accessible rgb-d cameras. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [58] B. Freedman, A. Shpunt, M. Machline, and Y. Arieli. Depth mapping using projected patterns, May 13 2010. US Patent App. 12/522,171.
- [59] A. Shpunt. Depth mapping using multi-beam illumination, January 28 2010. US Patent App. 12/522,172.
- [60] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846, Jan 1998.
- [61] Meinard Müller and Tido Röder. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '06*, pages 137–146, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association.
- [62] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.
- [63] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'81*, 1981.
- [64] Rizwan Chaudhry, Avinash Ravichandran, Gregory D. Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *CVPR*, pages 1932–1939, 2009.



- [65] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1998.
- [66] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.
- [67] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [68] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [69] *Instructing People for Training Gestural Interactive Systems*. ACM Conference on Computer-Human Interaction, 2012.
- [70] Chen Chen, Kui Liu, and Nasser Kehtarnavaz. Real-time human action recognition based on depth motion maps. *Journal of Real-Time Image Processing*, 2013.
- [71] F. Ofii, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 8–13, June 2012.