José Luís Machado de Figueiredo

Behavior Analysis in Autism Patients

Coimbra, February 2014

· U C ·

UNIVERSIDADE DE COIMBRA

Behavior Analysis in Autism Patients

Author

José Luís Machado de Figueiredo

Supervisor

Professor Doutor Paulo José Monteiro Peixoto

Thesis submitted in partial fulfillment
of the requirements to obtain the Master Degree in
Electrical and Computer Engineering

Jury

Professor Doutor Nuno Miguel Mendonça da Silva Gonçalves (President)
Professor Doutor Paulo José Monteiro Peixoto (Supervisor)
Professor Doutor Gabriel Falcão Paiva Fernandes (Vogal)

Integrated Master of Science in Electrical and Computer Engineering
Department of Electrical and Computer Engineering
Faculty of Science and Technology
University of Coimbra

Coimbra, February 2014

“Pedras no Caminho?

Guardo todas.

Um dia vou construir um Castelo…”

Fernando Pessoa.

Acknowledgements

I would like to thank my supervisor, Professor Doutor Paulo José Monteiro Peixoto who introduced me to this challenging and fascinating area *("No Pain no Gain")*.

I am especially thankful to my Mother, Brother, Aunt Maria, Grandparents, my Godfather, Aunt Manuela and Carlos Seixas who have always provided me with encouragement, advice, strength and unconditional Love.

I am thankful to Michéle Horta for allowing me to see the world with a constant smile on my face and for giving me the courage to pursue my dreams and turning them into reality.

To my brother Jaime Figueiredo, João Martins and João Santos who took their time and patience to record some of the video footage that made this thesis even possible.

To all my real friends that I had the luck to meet throughout my life and which always stood by me with kind words when I thought there was no way of solving a problem.

To my Computer Vision Laboratory colleagues and António Videira who helped me throughout this work and shared their enthusiasm about this project.

Last but not least a big thank you to my friends Rafael Fernandes and André Andrade. Without their help and incentive, each in their own way, this would have been much harder to achieve. Both of you know what I have been through this year. Thank you for your support, council, help and sometimes just your mere presence.

To all those who doubted me: Thank you for giving me the strength to prove you wrong.

Resumo

Esta obra insere-se no projecto de investigação *"Project Hometech"*. Este foca-se num sistema de detecção de anomalias que pode monitorizar e detectar padrões de comportamento de pacientes com autismo. O sistema possibilita a detecção de estereotipias gravadas separadamente em vídeos offline e a previsão de resposta de um paciente a um determinado estímulo. As estereotipias constam numa lista de comportamentos exibidos recentemente ou de actividades previamente executadas. Com os dados recolhidos e a informação processada, é possível que essa base de dados possa também ser utilizada para tratar outros pacientes com semelhantes incapacidades [1].

A doença de autismo pode manifestar-se de vários sintomas possíveis tais como o afastamento extremo, falta de interacção social, comportamento repetitivo e violento entre outros. O desenvolvimento de terapia comportamental é relativamente desafiante uma vez que cada paciente autista é um caso isolado devido à grande variedade de sintomas e casos de intensidade. Usando o conhecimento obtido e analisando os dados sobre as actividades diárias do paciente poderá revelar padrões que ligam essas actividades. Isso poderá proporcionar aos terapeutas algum conhecimento prévio de prováveis resultados comportamentais possíveis relacionados com as suas terapias.

Uma estrutura de reconhecimento de gestos foi criada com base num descritor local de movimento (LMD) com todas as informações necessárias [77]. A principal contribuição é propor um esquema de aprendizagem-classificação baseado em acompanhamento fiável através de pontos de interesse detectados utilizando o algoritmo Kanade-Lucas-Tomasi (KLT) em conjunto com o algoritmo mínimo de valores próprios desenvolvido por Shi-Tomasi [72]. Uma base de dados é gerada a partir do centroide, desvio padrão e velocidade média dos pontos de interesse acumulados. Na etapa final comparamos a base de dados gerada com uma sequência de testes com vários movimentos conhecidos e desconhecidos usando múltiplas máquinas de suporte de vector (SVM) binárias.

Palavras-chave: *Reconhecimento de Acção, Análise de Comportamento, Visão por Computador, Acompanhamento de Movimento, Detecção de Pontos de Interesse, Autismo, Treino, Classificação, Máquinas de Suporte de Vector, Aprendizagem por Máquina*.

Abstract

This work is part of a research project called *"Project Hometech"*. It focuses on an anomaly detection system, which can monitor and detect behavior patterns of autism patients. Using separately recorded behavioral patterns in offline video footage, the system could predict the response of a patient to a stimulus, given a list of recently displayed behaviors or completed activities. The knowledge thus gathered could also be used to treat other patients of similar disability [1].

The autism disease can manifest itself by a wide variety of possible symptoms such as extreme withdrawal, lack of social interaction, repetitive and violent behavior between others. The development of behavioral therapy is relatively challenging since every autistic patient is an isolated case because of the wide variety of symptoms and intensity case. Collecting that knowledge gathered regarding the patient and analyzing the data about a patient's daily activities could yield patterns linking these activities. This could thereby provide therapists with some foreknowledge of likely possible behavioral outcomes related to their therapies.

A gesture recognition framework was created based on a Local Motion Descriptor (LMD) with all the necessary information [77]. The main contribution is to propose a learning-classification scheme based on reliable tracking of detected features using Kanade-Lucas-Tomasi (KLT) algorithm combined with the Minimum Eigenvalue algorithm developed by Shi-described in [72]. A database is created using the centroids, standard deviation and mean velocity of the clustered moving points. In the final step, we compare and classify the generated database with a test sequence through known and unknown stereotyped movements using multiple binary Support Vector Machines (SVM).

Keywords: ***Action Recognition, Behavior Analysis, Computer Vision, Motion Tracking, Feature Detection, Autism, Training, Classification, Support Vector Machine, Machine Learning.***

Table of Contents

List of Tables

List of Figures

Nomenclature

Acronyms

ASD - Autistic Spectrum Disorders;

LMS – Local Motion Signatures;

HOG – Histogram of Oriented Gradient;

e.g. - Latin expression „exempli gratia „, which means “for example”;

i.e. - Latin expression „id est“, which means “that is,”;

SVM – Support Vector Machine

HMM – Hidden Markov Models;

CRF – Conditional Random Field;

aAM - Active Appearance models;

SVR - Support Vector Regression;

GPR – Gaussian Processes Regression;

PCA – Principal Component Analysis;

GB – Gigabyte;

LMS - Local Motion Signatures;

RAM – Random Access Memory;

CPU – Central Processing Unit;

PC – Personal Computer;

FPS – Frames per Second;

LMD - Local Motion Descriptor;

MMI - Maximization of Mutual Information;

RMS – Root Means Square;

pLSA - probabilistic Latent Semantic Analysis;

MKS - Microsoft Kinect Sensor;

RBS-R – Repetitive Scale of Behavior-Revised

Chapter 1

Introduction

Electronic devices have proven to be essential tools while becoming important in every person's daily routine. Nonetheless, these equipments are not fully autonomous because of their inability to learn and understand their surroundings and therefore they often need human interaction.

Over the last few years, several research groups in the areas of Artificial Intelligence, Image Processing, Machine Learning and Computer Vision, have been developing techniques that assist in understanding the ways in which humans interact with machines and each other.

One of the most active research areas in Computer Vision is visual analysis of human motion. Its goal is to track, detect and identify people correctly, as well as to interpret human behaviors from an image sequence. Motion analysis has caught a lot researcher's interest due to its potential for new applications in many areas. Recognizing human gestures and interactions is important for video annotations, automated surveillance and content-based video retrieval [4]. It can also be used for perceptual user interfaces, content-based image storage, video conferencing, athletic performance analysis and virtual reality [5].

The understanding of images and their source as computational intelligence has increased exponentially as computational power continued to grow year after year. New approaches surfaced as how to explore and extract further knowledge and information from images and video sequences. These were the first steps that led to the birth of Computer Vision and Machine Learning.

Computer Vision Technology allows the opportunity to study children´s behavior in a noninvasive way by providing means to capture behavioral data automatically and comprehend behavioral interactions between children and their caregivers and peers [3].

Machine Learning is a type of Artificial Intelligence where a system learns how to extract knowledge from data without specifically being programmed. This area focuses on computer program development that can teach itself expansion change when exposed to new data.

It is believed that modeling techniques and computational sensing can contribute significantly to the role of capturing, measuring, analyzing and understanding human behavior [3].

Very little is known about autism, its causes and its manifestations. The disease can manifest itself by a wide variety of possible symptoms such as extreme withdrawal, lack of social, repetitive and violent behavior between others. The development of behavioral therapy is relatively challenging since every autistic patient is an isolated case because of the wide variety of symptoms. Collecting that knowledge gathered about the patient and analyzing the data about a patient's daily activities could yield patterns linking these activities. This could thereby provide therapists with some foreknowledge of likely possible behavioral outcomes related to their therapies [1].

1.1 – Thesis Motivation

The aim of this thesis can be divided into three main objectives. The first one is the detection of people in an offline video sequence. For this to be possible, the Histogram of oriented Gradient (HOG) algorithm was used to detect the person's contours which later would allow the tracking to be more efficient [82].

The second objective is the implementation of a tracking algorithm which allows the person's feature points to be followed in every frame of a given video. The KLT feature point tracker has proven to be very effective and robust performing this task.

The last objective of this thesis is the detection of routine breaches in autism patients. This would allow them to be assisted immediately through surveillance cameras while not being attended and guarded by medical staff. A database file and video footage were filmed specially for this project so that the behavior detection could be performed using multiple binary Support vector Machine (SVM) Classification.

1.2 – State of the Art

There already exist a certain number of projects with the main goal of detecting stereotyped behaviors in an automatic way. The targeted number of stereotyped behaviors is usually limited and, usually, even reduced to a single one (normally hand-flapping).

The number of instructed actors who engage in those identical movements similar to observed stereotypes of Autistic Spectrum Disorders (ASD) are reduced.

The possibility of creating a computer generated system that can automatically identify risk situations and stereotypes can prove to be a valuable asset. This allows the ability to collect huge amounts of information without the need of substantial human effort. Automatic activity recognition usually focuses on just one subject, describing it in terms of predefined actions. James Rehg [3] proposed such an automated system for stereotype identification which can help the correct detection in early stages of ASD in children of lower age. For that to be possible, interaction patterns between individuals have to be identified as well as individual actions in an isolated way. Hashemi [7] and his team used Computer Vision and Machine Learning based technology and proposed a new method to extract behaviors from video sequences to identify ASD in an early stage in children. The developed system analyses the position and head movement, torso, arms and legs of each child trying to recognize restrictive patterns and repetitive behaviors that are suggestive of autism. A big problem with these approaches is that they are conducted offline and not in real-time. The aim of Project Hometech is that the abnormal behavior detection should be identified in real-time so that alarm signs can be sent to the clinical staff for eventual crisis situations.

The following sections will introduce the existing techniques of the state of the Art. In section 1.2.1 we discuss the topic regarding Automatic Analysis of human behavior and facial expression detection and its subtopics automatic human detection, automatic understanding of human behavior, facial expression detection and head pose estimation in 2D.The following section 1.2.2 focuses on the topic called stereotyped behavior identification.

1.2.1 – Automatic Analysis of Human Behavior and Facial Expression Detection

Other known techniques are in the domain of Automatic Analysis of human behavior and facial expression detection. This area has been flourishing as research topic due to its enormous potential and application fields. The actions of one or more individuals are analyzed from arbitrarily placed cameras over the scene. The camera placement does usually not follow rigidly established rules [8]. This implies the

existence of video image analysis methods and algorithms which are invariant to the cameras vision angle. Unfortunately, the majority of the proposed methods for human behavior detection are susceptible to some form of restriction regarding the camera point of view. For example, the face has to point towards the camera or parallel to the image plan [9, 10, 11, 12, 13]. These type of limitations due to camera point of view dependency are difficult, if not impossible, to recreate in realistic scenarios. Various researchers have pointed out just how important the camera perspective is towards the objects of interest in the overall performance of the existing methods for human motion analysis [14, 15].

The most recent developed work in human motion analysis can be divided in two categories: Pose Estimation and Action Recognition. The difference between these two resides in the fact that the first one prioritizes the 3D human pose estimation problem obtained from individual images from image sequences. The second category focuses on how to deduce and understand human actions from identified patterns detected in Images. These two methods are directly related in the sense that 3D pose estimation can posteriorly be used as a starting point for various action recognition methods.

Pose estimation methods can be divided into the following categories: Pose representation based on 3D models, Pose representation without using models and Pose representation based on [16, 17, 18, 19]. This type of approximation involves some computational complexity in the estimation models and while updating it during the action. This normally implies elevated computational costs, leading to the near impossibility of a real-time processing scenario.

Action recognition can be represented in two categories as well: Template based classification and State Space methods. Template based approximation is based on two phases: The ways of representing actions based on predefined characteristics existing in the image have to be identified in the first part. Action recognition can be considered a normal classification problem in pattern recognition in the second phase [20, 21]. State Space approximations define each static pose as a state, resorting to a probabilistic model to generate mutual connections between several of these states. Any sequence of movement can be considered as a route through several states of these static positions [22, 23].

1.2.1.1 – Automatic Human Detection

Automatic detection of humans in image sequences seeks to distinguish moving people from de scene background. This is fundamental for the detection of human behaviors since the pose detection and action recognition are usually very dependent on the system's performance regarding human detection [5, 24, 25]. It is fundamental to have a good segmentation of the pretended targets on the scene so that human action analysis has an elevated performance. This can be achieved with movement segmentation or with object classification.

Movement segmentation is used to detect regions in the image which correspond to moving objects and that can potentially be correlated with human beings present in the scene. The objective here is to detect image zones that can be of interest for tracking algorithms and activity analysis [24]. Different zones moving in an image can correspond to different moving objects in the scene. It becomes necessary to classify each moving object so that potential human beings can be identified and separated from other moving objects. The scale factor can influence these results.

There are two categories for object classification: Form based classification and movement based classification. Form based classifications pretend to identify form characteristics of each moving object. The problem is posteriorly considered as a pattern recognition task [26, 27, 28, 29, 30, 31]. Human body articulations and the different perspectives between cameras create an enormous number of possible body appearances. This makes it hard to distinguish a human body from other moving objects using merely the form. Movement based classification uses intrinsic periodic properties of articulated body movement to distinguish human beings from other moving objects. Auto-similarity is periodic and allows methods of temporal frequency domain analysis to be used to distinguish periodic movement [32].

1.2.1.2 – Automatic Understanding of Human Behavior

Automatic understanding of human behavior analyzes and recognizes motion patterns. It tries to create high level descriptions of conducted human actions and their interactions in several application scenarios. Behavior understanding can be divided into two areas: Action Recognition and Behavior Description. One major problem of action

recognition is the human action speed since its representation should be independent. Other flaw may be the movement concurrence of individual body parts since they can move with concurrent periodic synchronization. Another problem can be the variability of human movements. The same action can be done several times by the same individual or by several people. They can demonstrate variations which makes the movement identification a complex task. The movement comparison with the previously existing system knowledge can turn out to be quite difficult to have an exact match regarding a particular movement.

Template Matched Methods converts an image sequence into static patterns whether they are spatial or temporal. It is hoped that these patterns are capable of discerning the expected recognizable actions. These patterns are then compared with a priori obtained similar representations [33]. The advantage of this method is the low computational cost and simple implementation. In contrast, they are more susceptible to noise and to temporal variation of actions. Bobick and Davis [34] were the first to suggest the use of temporal templates to represent human actions. Hu moments were used as discriminant action characteristics. After this representation, action recognition started to be considered as a conventional classification problem in pattern recognition

State Space based approximations define each static pose as a state using a probabilistic model to generate mutual connections between several of these states. These methods normally use the recovered results of the tridimensional pose so that they can build and identify representation forms for actions. Hidden Markov Models (HMM's) [35] are a sophisticated technique for the analysis of time variant systems. They have been used to describe temporal relations inherent to human actions [36, 37, 38, 39]. [40] decomposed a space of elevated dimension corresponding to the set of joints on the articulated model used to represent a human pose in a set of feature spaces. In those, each feature corresponds to the movement of each joint or the combination of those related joints to each other by the same type of movement. Given an Image Sequence during the recognition process it is possible to calculate the probability of observation simply using individual HMMs to recognize each class of shares. On the other hand, a scheme based on AdaBoost can be used to detect and recognize each feature in the image.

Although this area has evolved a lot in recent years, it still has to solve some well-known problems. The use of monocular sequences remains problematic for estimation. The combination of methods based on examples and model-based tracking

has proven to be a promising direction to solve this obstacle [41, 42].It is clear that inferring the 3D pose using example based methods makes it quite complicated due to the high number of parameters that are required to estimate, not to mention the ambiguities introduced by the perspective projection. The state of the art suggests contextual restrictions and the usage of fusion between several image feature extraction based methods could be viable alternatives to solve the problem [40, 43]. It is useful to achieve an acceptable compromise between the computational cost and performance recognition methods in approximations based on state spaces. Conditional random fields (CRF) have been identified as promising when compared with the HMMs in the analysis of human motions [43]. The automatic comprehension of human behavior is a complex task since the same behavior can have different meanings depending on the spatial and behavioral context in which the task is performed. The investigation of behavior patterns through self-organization and self-learning applied to unknown scenes is a chance to explore.

1.2.1.3 – Facial Expression Detection

Facial Expression Detection is another very important topic. Head pose estimation is a process that can deduce its orientation from video images. However, numerous intermediate steps are necessary to obtain this estimation and to transform a head based representation on image points into a directional concept. An ideal face position estimator must demonstrate invariance to a set of factors that can alter or otherwise distort the perception of the head in an image. These factors include phenomena such as the geometric distortion that may be caused by the camera due to poor calibration of its own, multiple light sources to change the brightness of the environment and on surfaces on which incur complex and non-static image backgrounds. This could increase the image complexity and partial or complete occlusions of interest elements and thereby obstructing the visibility of what is pretended to be analyzed. Some variations exist even on a biological level such as morphological characteristics of each individual as skin tone, beard, hair, eyebrows, facial expressions which can distort the face appearance and even some props such as glasses and hats. A pose estimation algorithm should be robust and able to deal with some of these mentioned factors since its first step consists in face detection. This task

is of particular importance since it will be responsible for achieving relative information regarding the head which will be used in consequent tasks for pose estimation.

Various face detection techniques are proposed for applications in image intensity (grayscale) or full scale (color) [44]. These techniques can be grouped into four categories: Knowledge-based methods, Feature Invariant Approaches, Template Matching Methods and Appearance-based Methods.

Knowledge-based methods incorporate attribute-based knowledge on the human face, capturing the relationship between those attributes [45, 46].

Feature Invariant Approaches attempt to track facial features even when the face is viewed from alternative points of view or when it is subject to lightning variations. All of this with the purpose of face detection [47, 48].

Template Matching methods store several faces or several facial features for posterior correlation acquisition between an image that is to be analyzed and the previously stored patterns [49, 50].

Appearance-based methods are similar to template models since both use a database, but instead of using this information for direct comparison they use it as a training set to capture representative information of human faces [51, 52, 53].

1.2.1.3.1 – 2D Head Pose Estimation

Several approaches can be found in the literature for head pose estimation. They can be divided into Appearance Template methods [54, 55], Classification based methods [56, 57], Regression based methods [58, 59, 60], Manifold Embedded based methods [61, 62, 63, 64, 65, 66], Active Appearance models (AAMs) [67, 68] and Geometric methods [69, 70].

Models based on appearance and classification are based on comparing trained face images which are pre-cataloged in a set of discrete angles with test images for which it is intended to determine the head position. Typically these methods are sensitive to uneven sampling whether in the training or test images. Training images must show uniformity in the same image set or otherwise the subsequent comparison with test images will induce incorrect results.

Models based on regression estimate head poses, generating linear or nonlinear functions, between images of continuous angles. Several regression methods are possible where some of the most common are Support Vector Regression (SVR's) and Gaussian Processes Regression (GPR) processes. These models are similar to the classification methods but are commonly used in continuous angle measurements and they suffer from the same problem of nonuniform samples in the training and testing phases.

Models based on embedded techniques produce low-dimensional representations of facial features such as the Principal Component Analysis (PCA) [71] which is categorized as a technique of dimensional reduction. The vast majority of embedded methods are unsupervised. That means that they are unable to extract feature information which is necessary for assigning a class of angles. There is no guarantee that the characteristics of the subject such as age and facial expressions do not overlap with features relating to pose. Examples of success of these methodologies are models based on Eigenfaces [64, 66], models that use Isomaps Embedding [61, 62], and Laplacian Eigenmap Embedding [65]. However these models still rely on classification and regression techniques to estimate the angle and they still possess the same above-mentioned disadvantages.

1.2.2 – Stereotyped Behaviors Identification

The identification of stereotyped behaviors and its framework are vital topics. One of the characteristics that defines ASD is restrictive and repetitive behaviors. Although it is not a precise definition they correspond to motor movements and are often described as rhythmic, involuntary and purposeless.

The detection of stereotypies is an opportunity to redirect these behaviors to adaptive activities. The use of tools based on computer vision can bring several advantages like permanent monitoring at a reduced operating cost, objective evaluation of the effectiveness of therapies by analyzing the differences in frequency and duration of stereotypies between the intervention groups and the control groups. Other benefits are the aid in early detection of potential cases of autism and its potential as a diagnostic tool, the opportunity for automatic intervention strategies based on Information

Technology tools such as redirection through distracting stimuli and the automatic detection of relevant behaviors from clinical point of view.

1.3 – Used Resources

This project was developed partially on a Notebook with 4 GB of RAM powered by an Intel Core 2 Duo CPU clocked at 2.27 GHz. Most part of the tests were executed on a desktop PC with 8 GB of RAM and an Intel Core i7-870 Processor clocked at 2.93 GHz.

The operative system used on the Notebook was a dual boot configuration with Windows 7 at 64 bits and Ubuntu LTS 10.04 equally at 64 bits. The desktop PC worked solely under Ubuntu LTS 10.04 at 64 bits. All tests and programming were performed in Linux environment.

In terms of software, OpenCV was considered at first and tested on during some time. After careful analysis and consideration, a switch was made to MathWorks MATLAB. It provided better algorithm development tools and helped to simplify the analysis and preview of information. The Image Processing was developed using some tools available in the *Image Processing Toolbox* and some of the video analysis and algorithms were designed seizing the *Computer Vision Toolbox*. These Toolboxes of the mentioned MathWorks software allow computers with multiple cores to use their full potential.

The Video Footage was obtained using a Sony DCR-DVD 106 DVD cam at standard 720x576 resolution at 25FPS and afterwards edited using Avidemux software. During the development of this project, new footage was recorded with the 16 GB version of a black Apple iPhone 5. Video Footage was recorded in 1080p and then resized using the mentioned software to 640x360 at 24 FPS to maintain the aspect ratio.

.

1.4 – Thesis Structure

The next few chapters will discuss in detail the work developed to achieve the proposed objectives.Chapter 2 provides information about the theoretical concepts used to support the choices taken during the development of this project. Brief descriptions of practical elaboration steps that have been considered throughout the project,

including abandoned approaches because of its negative implications, are also included. Chapter 3 presents the results of the numerous taken approaches that have been considered and the information to support those decisions are demonstrated. Chapter 4 presents a discussion of the results and some ideas for future work.

Chapter 2

Theory of Support

2.1 – Summary

This chapter will introduce the used methodology to meet the stipulated objectives. The videos for this project were all recorded on purpose with the previously described characteristics and materials. It had the objective to automatically and reliably identify and categorize a selected stereotype behavior.

The proposed action recognition approach consists of people and feature detection in an offline tracking module and an action recognition module. The detected interest points, known as spatio-temporal features, would later be stored into a local motion descriptor (LMD) database.

To reduce computational effort, people detection is performed only in the first video frame, using the well-known Histogram of oriented Gradient (HOG) algorithm [73].

The person is only detected once and then the Kanade-Lucas-Tomasi (KLT) algorithm, combined with the Minimum Eigenvalue algorithm developed by Shi-Tomasi [72], tracks the feature points across the video frames. Once the detection locates the person, the next step is to identify feature points that can be reliably tracked. Various algorithms were compared for the tracking procedure.

The obtained feature points were then grouped in 2 different clusters using K-means algorithm: *"moving"* and *"not moving"* points. The information extracted from the clusters was stored into the database for posterior classification.

Finally, multiple one-against-all binary Support Vector Machine (SVM) classifiers were trained to classify the different stereotype behavior classes [74]. Figure 2.1 describes the whole process.

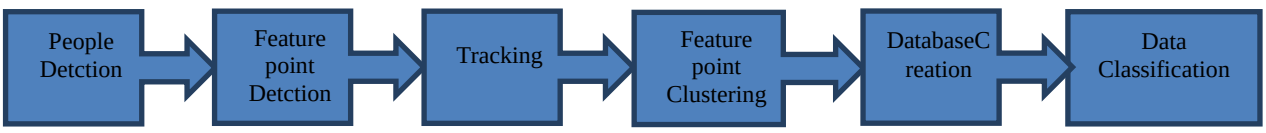

Figure 2.1: Stereotype recognition procedures.

2.2 – People Detection

Reducing the Region of Interest (ROI) for Feature Detection to the ROI of a detected person has several advantages. It helps to significantly reduce the computational effort in finding unnecessary features in the image. All the coordinates were relatively to the detected ROI area and not regarding the Image coordinates. This eliminates any possible conflict regarding a dependency of Image coordinates for future stereotype behavior recognition. This was accomplished using the HOG algorithm described in [73]. This method is known for its good results for pedestrian detection [82]. Figure 2.2 exemplifies the procedure.

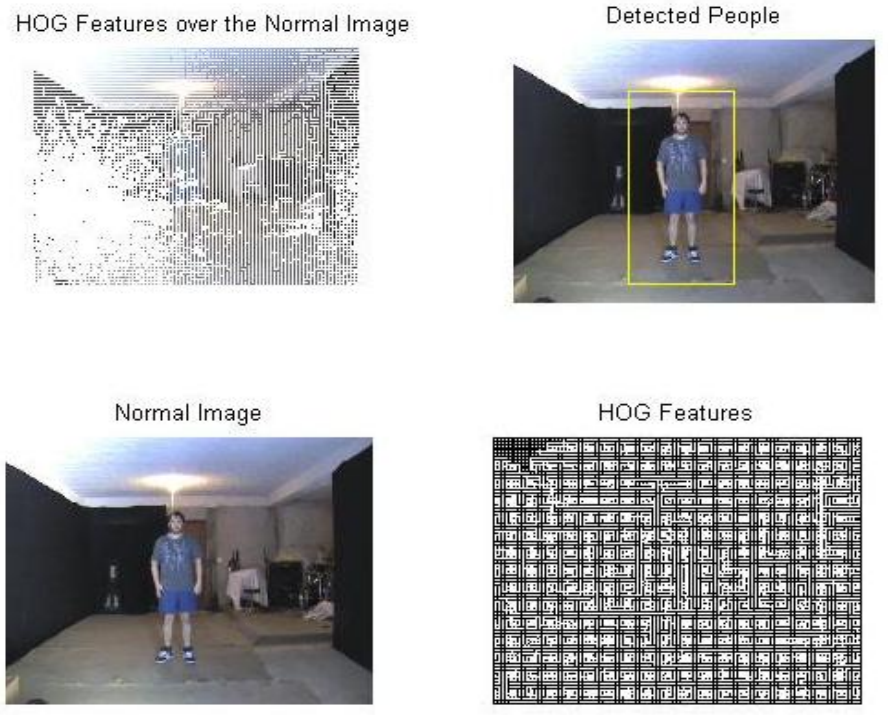

Figure 2.2: People detection procedure.

2.3 – Feature Point Detection

Feature Detection algorithms such as SURF [86], Harris [85], FAST [84], MSER [87] and Minimum Eigenvalue [72] were the ones that most stood out in the existing literature. The following figures 2.3 to 2.7 demonstrate the performance of each of the mentioned algorithms.

Figure 2.3: FAST feature performance [84].

Figure 2.4: Harris feature performance [85].

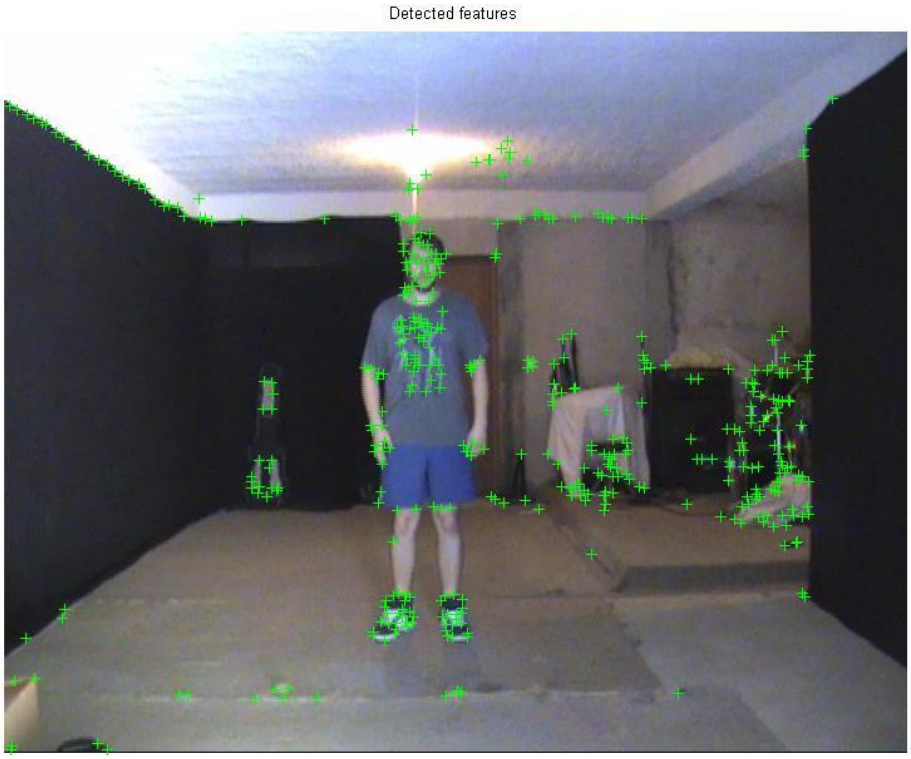

Figure 2.5: Minimum Eigenvalue feature performance [72].

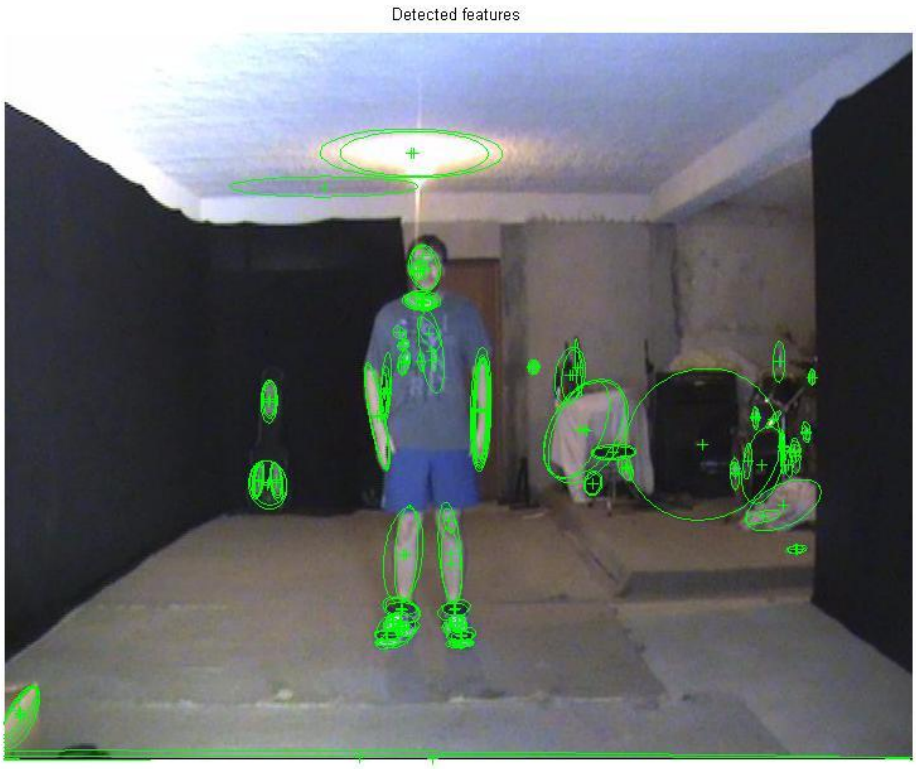

Figure 2.6: MSER feature performance [87].

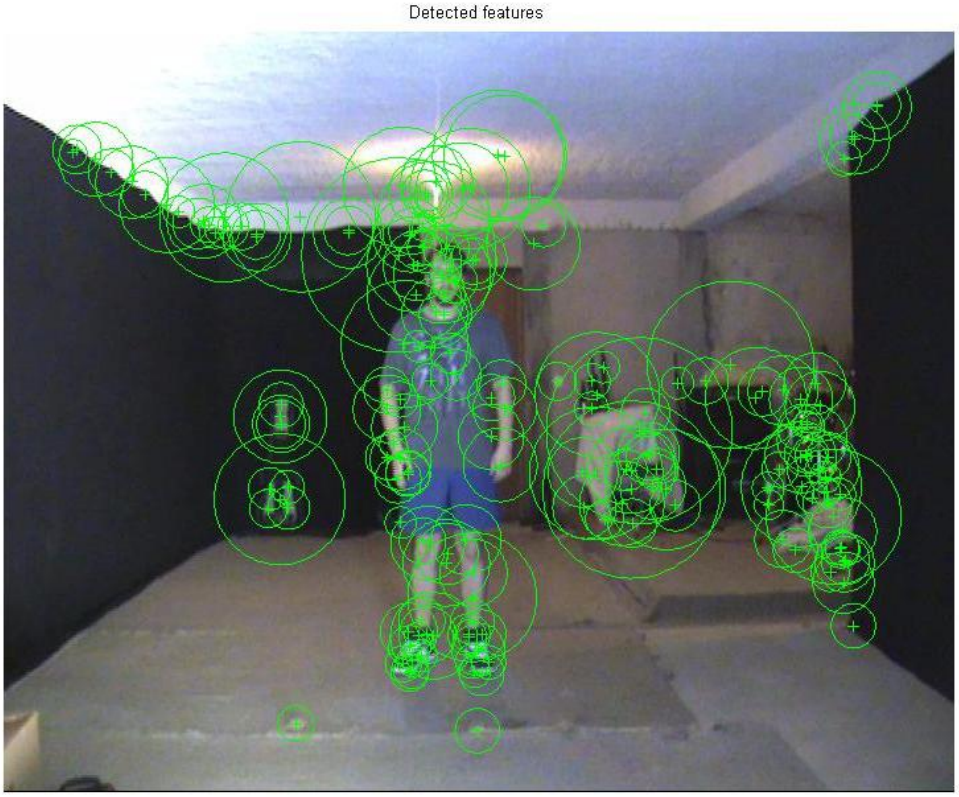

Figure 2.7: SURF feature performance [86].

The best feature detection algorithm fitted for this project's purpose was the minimum eigenvalue algorithm developed by Shi-Thomasi [72]**.** Given an image I in gray-scale, corresponding to a moving region, we first compute gradients g_x along the x axis and g_y along the y axis. Then, simple filters are applied as $[-1 \quad 0 \quad 1]$ for g_x and $[-1 \quad 0 \quad 1]^t$ for g_y. In [73] the author justifies this choice since it demonstrates that these filters consume less processing time than other gradient based features and are just as efficient as any another (e.g. Sobel operator, Derivative of Gaussian, etc). Then, for each pixel p in the image and in a window (u, v) centered on the considered pixel, we compute the 2x2 Hessian matrix defined by the equation 2.1.

$$H_p = \sum_u \sum_v \begin{bmatrix} g_x^2(u,v) & g_x g_y(u,v) \\ g_x g_y(u,v) & g_y^2(u,v) \end{bmatrix} \quad (2.1)$$

After that, we calculate the eigenvalues λ_1 and λ_2 of the Hessian Matrix H_p, i.e. the roots of $\det(H_p - \lambda \mathrm{I}) = 0$ [72] prove that $\min(\lambda_1, \lambda_2)$ is a better measure of the corner strength than the one given by the Harris corner detector [77].

2.4 – Tracking

As previously mentioned, KLT algorithm was sued for the tracking procedure. This can be divided into three parts: Detect a person, identify features to track and track the features over the frames. Kalman filter tracking was considered and tested but abandoned because some stereotyped movements led to inaccurate predictions. For a matter of comparison, the number of feature points per frame was established as 2000. Once the detection locates a person, the next step in the example identifies feature points that can be reliably tracked. This project uses the previously mentioned minimum eigenvalue algorithm described in [72]. It monitors the image features quality during tracking by using a measure of feature dissimilarity that quantifies the appearance change of a feature in between the first and current frame. The idea is fairly simple since the dissimilarity is the feature's RMS residue between the first and the current frame.

In some test videos throughout this project it was noticed that dissimilarity grows too large and therefore the point should be abandoned [72]. In those cases whenever a point is abandoned a new one is generated in order to keep the desired number of features good track. Those new points were generated using again the Shi-Tomasi algorithm and selected with a random sample to guarantee randomness. This was done to avoid any matrix dimension mismatches for the future creation of the LMD from the tracked point. The following figure 2.8 exemplifies the tracking steps.

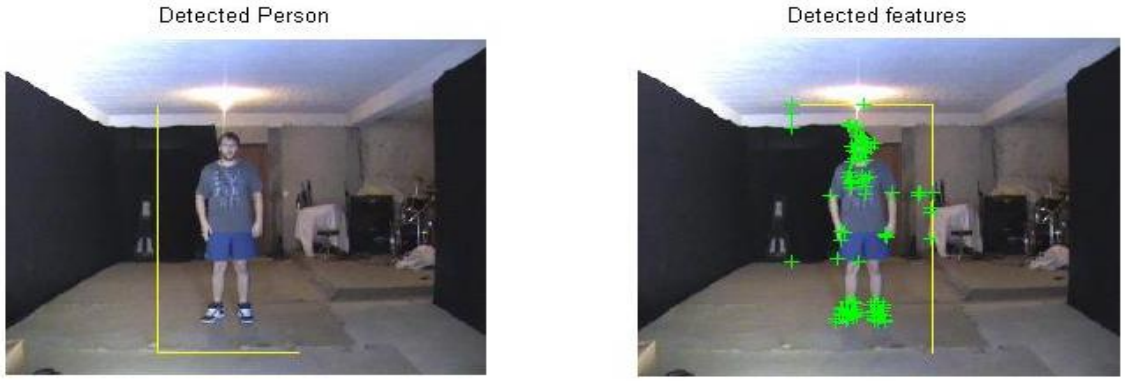

Figure 2.8: People tracking procedure.

The following figure 2.9 explains how new points are generated

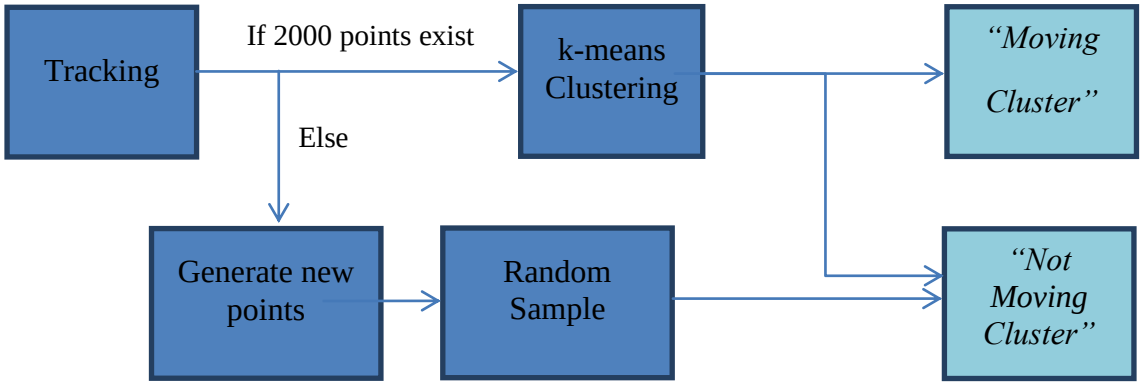

Figure 2.9: Generation of new points procedure.

The random points are grouped directly into the "*Not Moving*" cluster since it has no previous velocity history.

2.5 – Feature Point Clustering

The previously obtained feature points were grouped in 2 clusters according to their velocity using the K-means algorithm [88] which is a well-known unsupervised learning algorithm that allows the solution of the clustering problem. This allowed to significantly reduce the number of feature input to be analyzed by the classifier. This algorithm is easily implemented and is computationally efficient. It allows the processing of very large number of samples. Possible applications include methods for

similarity grouping, nonlinear prediction, approximating multivariate distributions, and nonparametric tests for independence among several variables [81].

This algorithm aims at minimizing an objective function, in this case a squared error function. The objective function is given by (2.2).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (2.2)$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j is an indicator of the distance of the $n$data points from their respective cluster centers. The algorithm is composed of the following steps:

1. *Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.*
2. *Assign each object to the group that has the closest centroid.*
3. *When all objects have been assigned, recalculate the positions of the K centroids.*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*

K-means is a simple algorithm [88] that has been adapted to many problem domains but it also has its weaknesses. It does not necessarily find the best fitted configuration, it is very sensitive to the initial randomly selected cluster centers besides the necessity to know the number of clusters before its action. It was run during 5 iterations to reduce this inconvenience. Figure 2.10 shows an example of the clustering of the points into the two previously mentioned clusters. Red color shows the *"Moving"* cluster while blue color represents the *"Not Moving"* cluster.

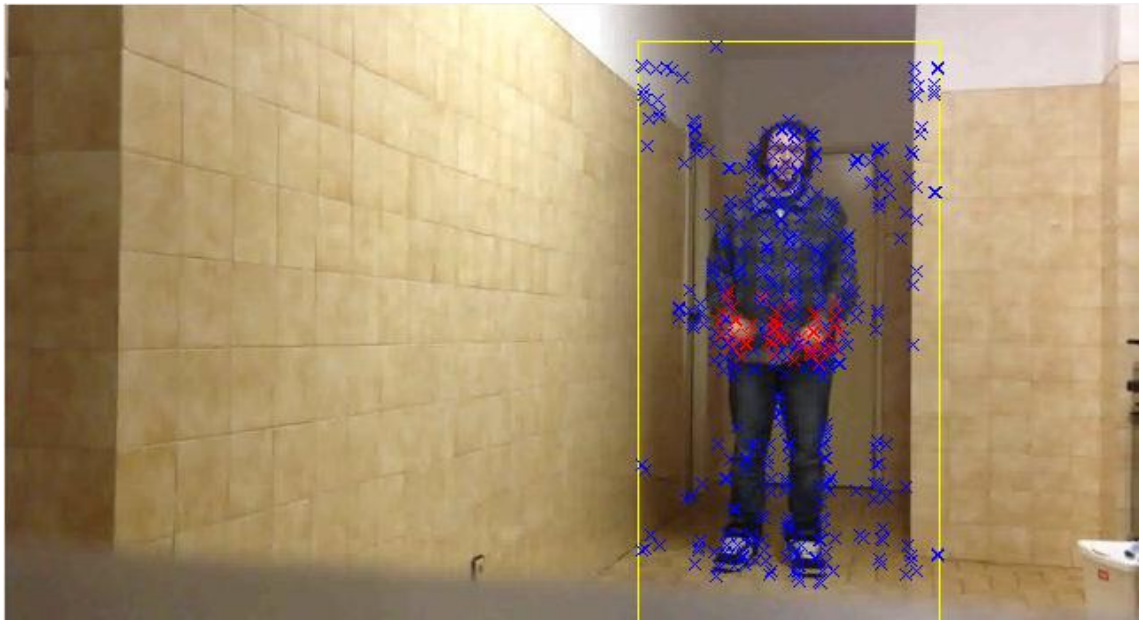

Figure 2.10: Feature point clustering.

2.6 – Database Creation

During the tracking procedure using the KLT algorithm, all the positions in x and y coordinates of the feature points, as well as the velocity and acceleration vector coordinates, were stored into a spatio-temporal feature vector, the LMD, as exemplified by equation 2.3.

$$LMD = [P_x \quad P_y \quad V_x \quad V_y \quad A_x \quad A_y] \text{ (2.3)}$$

Where P_x and P_y are regarding the position, V_x and V_y relatively to the velocity and A_x and A_y towards the acceleration.

The huge number of points to classify could lead to excessive computational time to train the SVM and to classify the data. Therefore, a new approach was initiated using [3] as inspiration.

The feature points were simply clustered in two categories*: "Moving"* and *"Not Moving"*. The mere analysis of the centroids and the standard deviation points belonging to the moving cluster proved to be a simpler and more elegant way to describe the stereotype movements. The LMD used can be demonstrated by (2.4).

$$Database = [Centroid(\mathrm{M})_x \quad Centroid(\mathrm{M})_y \quad std(\mathrm{M})_x \quad std(\mathrm{M})_y \quad V_x \quad V_y] \tag{2.4}$$

Where $Centroid("M")$ represents the mean value of the position of the points belonging to the moving cluster $"M"$ in x and y coordinates, in each frame. The mean value can be obtained using (2.5).

$$mean = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.5}$$

$std("M")$ represents the standard deviation of the position of the points belonging to the moving cluster $"M"$ in x and y coordinates in each frame. The standard deviation value can be calculated using expression 2.6.

$$std = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{1}{2}} \tag{2.6}$$

V_x and V_y represent the mean velocity of the points belonging to the moving cluster $"M"$ in x and y coordinates, in each frame. Those were later stored into the Database and saved into a MathWorks mat file and used to train the SVM classifier.

2.7 – Data Classification

After a careful analysis most of the available classifiers commonly used in this kind of project, it was decided to use multiple one-against-all binary SVM Classifiers. It maps the data into a richer feature space, including nonlinear features and constructs a hyper plane afterwards in that space so that all other equations are equal.

According to [74] the performance of the multiple linear SVM is superior to the competing algorithms. In case of stereotype movements and their actions where intra-class variability is significant, the supervised SVM classifier algorithm should perform better. The work developed in [74] demonstrates that the multiple binary SVM classifiers outperform generative pLSA or Naïve Bayes classifiers. This is both due to

over-fitting, as well as the poor performance of the unsupervised methods in segmenting the codebooks into distinct action classes. This is due to the large intra-class variability in visual appearance of human actions.

To classify all the four stereotype movement categories, we adopt 4 binary SVM classifiers [76], so that video instances associated to the respective stereotype behavior class is within the same and the other test video is in another. The binary SVM classifier decides whether a stereotype movement video belongs to the *"A"* or *"not A"* stereotype movement class [74]. Given N labeled training data

$$\{x_i, y_i\}, i, \dots N;$$

$$y_i \in \{0,1\}, x_i \in R^d$$

where x_i represents the distribution of the spatio-temporal interest points for each video i with d dimensions and y_i is the binary action movement label. The SVM classifier's intent is to find an optimal hyper-plane

$$w^{T-}x + b = 0$$

between the positive and negative samples. We assume that there is no prior knowledge about the distribution of the action class videos. We use a conventional learning approach for SVM which seeks to optimize the following problem shown by (2.7).

$$min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_i \xi_i \quad (2.7)$$

is subject to :

$$y_i(w.x_i + b) \geq 1 - \xi_i \text{ for } i, \dots N$$
$$\xi_i \geq 0 \text{ for } i, \dots N$$

Where C is an adjustable parameter that represents the penalty constant which is determined by cross-validation. For each action classifier, the classification decision score is stored. Among all classifiers, the one which results in the highest classification score is chosen as the stereotype behavior class and the outcome of each video's classification is labeled accordingly.

Chapter 3

Experimental Results

3.1 – Summary

This chapter will expose the obtained results from all the previously discussed methods and techniques, as well as data from conducted tests. The objective is to test the undertaken decisions and analyze the robustness of the developed algorithms. Various tests were conducted using different setups.

The Training Video Sequence consisted of four well known stereotype movements described in Table 3.1. Those were based on the Repetitive Scale of Behavior-Revised (RBS-R). More on that can be found in the Appendix A.

Name of the Stereotype	**Description of the Stereotype Movement**	**Stereotype Name in the Training Tequence**
"Body Rocking"	Crossing your hands over your chest, putting your hands on your shoulders and nodding	*Mov01*
"Nodding"	Heads nods up and down	*Mov02*
"Shake Head"	Shaking head to the left and right	*Mov03*
"Hand Flapping"	Flaps hands like a bird	*Mov04*

Table 3.1: Stereotype movements present in the training sequence.

As an extra movement, to merely demonstrate that the results were coherent, hand waving was added in the test sequence to see if it could be successfully detected and distinguished from all the others.

3.2 – Histogram Analysis

As mentioned previously, the method in [3] was used as guideline. The analysis of the histogram of each training stereotype movement clearly shows that each one has a different and characteristic histogram. This alone would be helpful to classify the movements. The data input for the SVM Classifier should be normalized. The following figures 3.1 to 3.4 prove that each histogram has its own distinct histogram regarding the already mentioned mean, standard deviation and mean velocity values of the points belonging to the moving cluster "*M*" in x and y coordinates, in each frame.

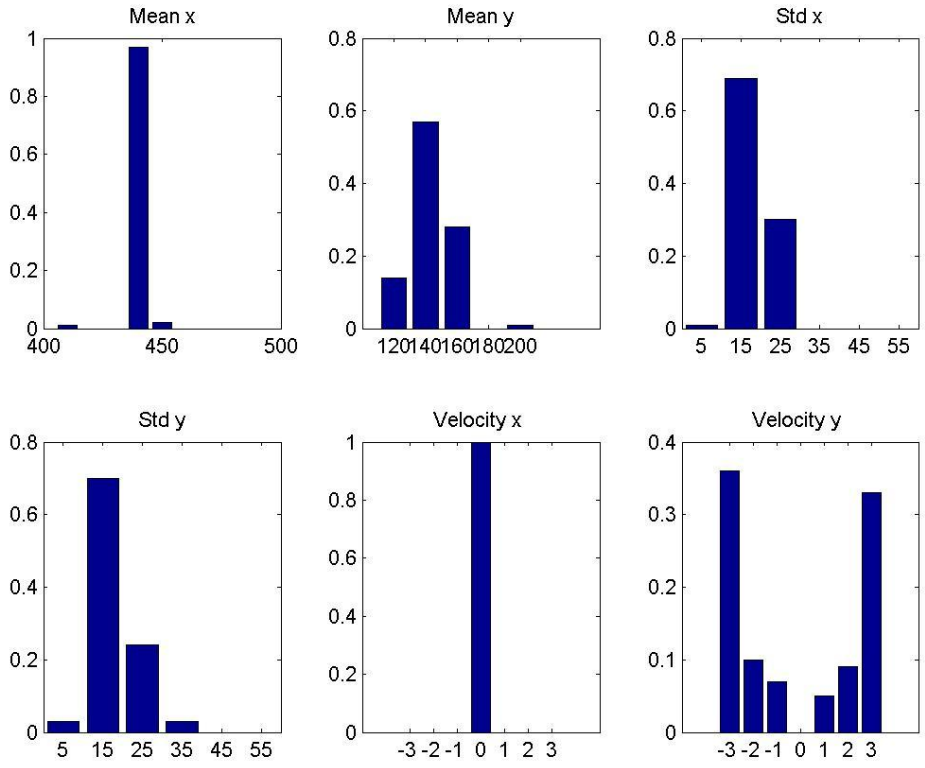

Figure 3.1: Histogram of the stereotype movement *"Body Rocking"*.

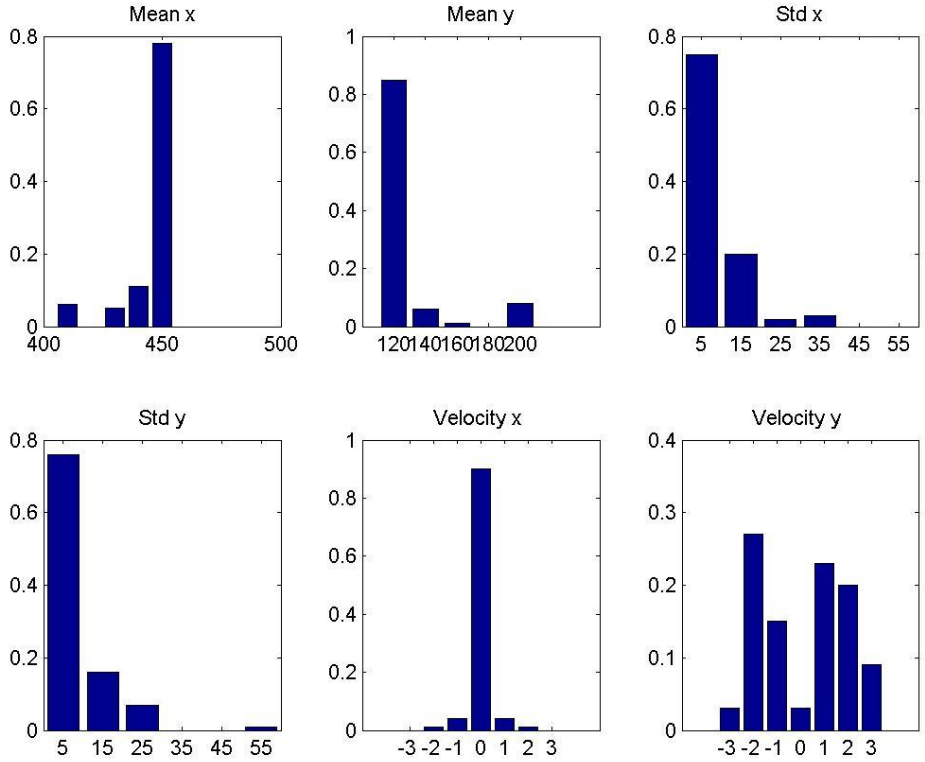

Figure 3.2: Histogram of the stereotype movement "*Nodding"*.

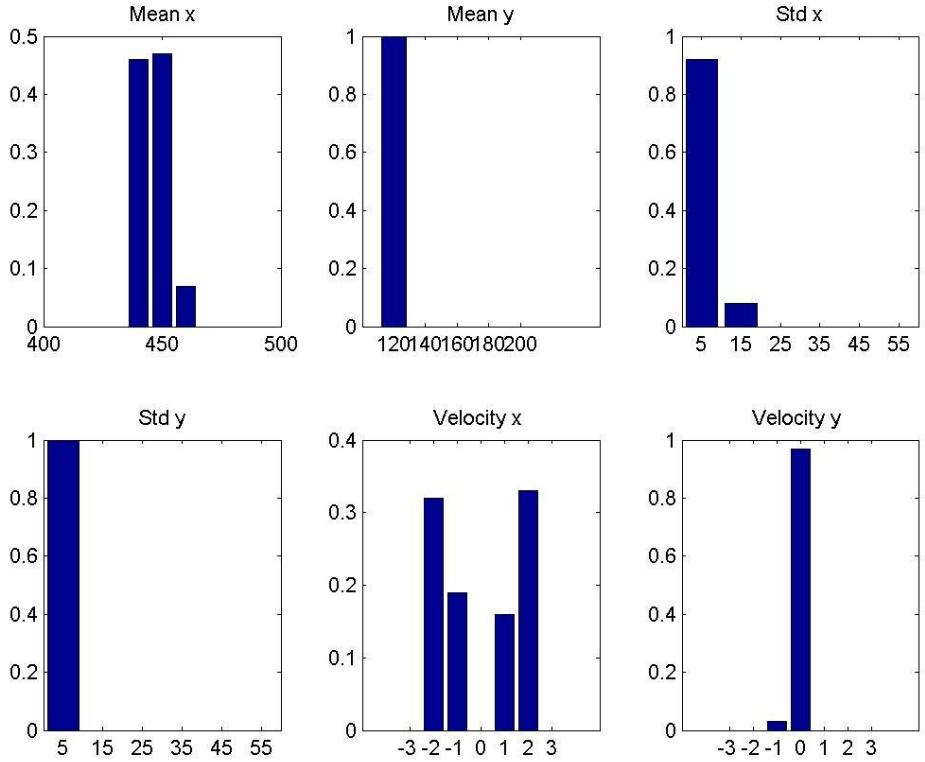

Figure 3.3: Histogram of the stereotype movement *"Shake Head"*.

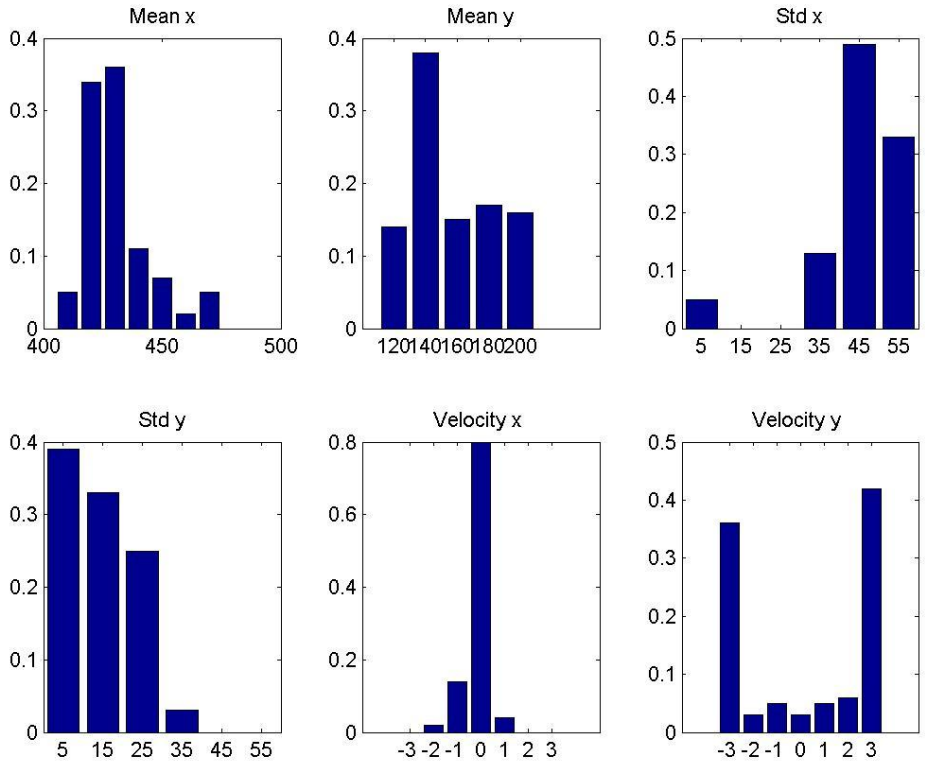

Figure 3.4: Histogram of the stereotype movement *"Hand Flapping"*.

Since the main interest of this project is the detection of spatio-temporal relations we cannot use the histograms for sole purpose of classification. They have the disadvantage to hide the order in which the movements were executed and that would eliminate the chances of successful classification of a stereotyped behavior.

3.3 – Classification

As previously mentioned, a multiple binary SVM Classifier was used to classify the stereotyped behaviors. It is a discriminative machine learning algorithm based on the structural risk minimization induction principle [75]. Each classifier distinguishes only between *"Mov"* or *"Not"* for its respective movement. The training footage used were frame shifted videos due to shortage of material. Therefore the classification process could prove to be inaccurate or show limited results. The minimum duration of a movement was considered to be 100 frames. Every binary SVM Classifier was trained with videos of the other movements as negative training samples. Table 3.2 shows the

order of the executed stereotype movements present in the Testing Sequence. It was elaborated in a different environment, clothes and had several movements to distinguish it from the static testing sequence.

Stereotyped Movement	**Description of the Stereotype Movement**	**Stereotype Name in the Training Sequence**	**Sequence Length(Frames)**
"Body Rocking"	Crossing your hands over your chest, putting your hands on your shoulders and nodding	*Mov01*	*400*
"Nodding"	Heads nods up and down	*Mov02*	335
"Wave hand"	Waving Hand repeatedly	*Unknown*	*255*
"Hand Flapping"	Flaps hands like a bird	*Mov04*	*295*
"Nothing"	No movement	*not*	*137*

Table 3.2: Stereotype movements present in the testing sequence.

Although there are some slight variations, it can be seen that they have some similarities nonetheless. Those variations can be due to scaling problems, velocity of the stereotype execution, different clothing and due to some tracking issues which were referred in the state of the art. Those comparison results can be seen in figures 3.5 to 3.12. They show the histograms of the test and training sequence.

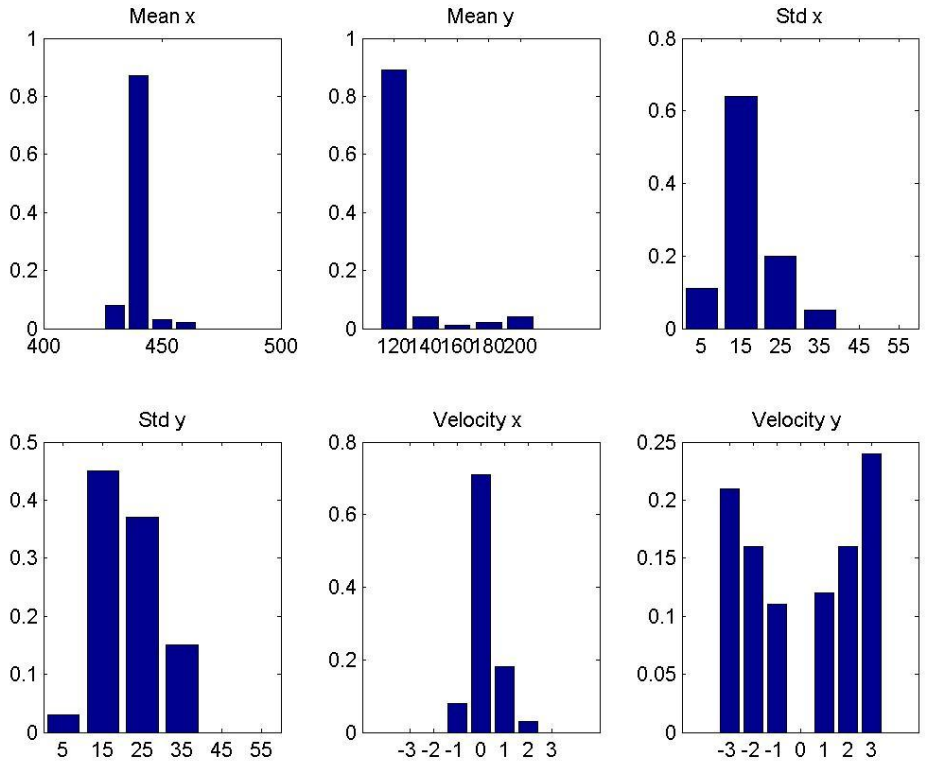

Figure 3.5: Test histogram of stereotype movement *"Body Rocking"*

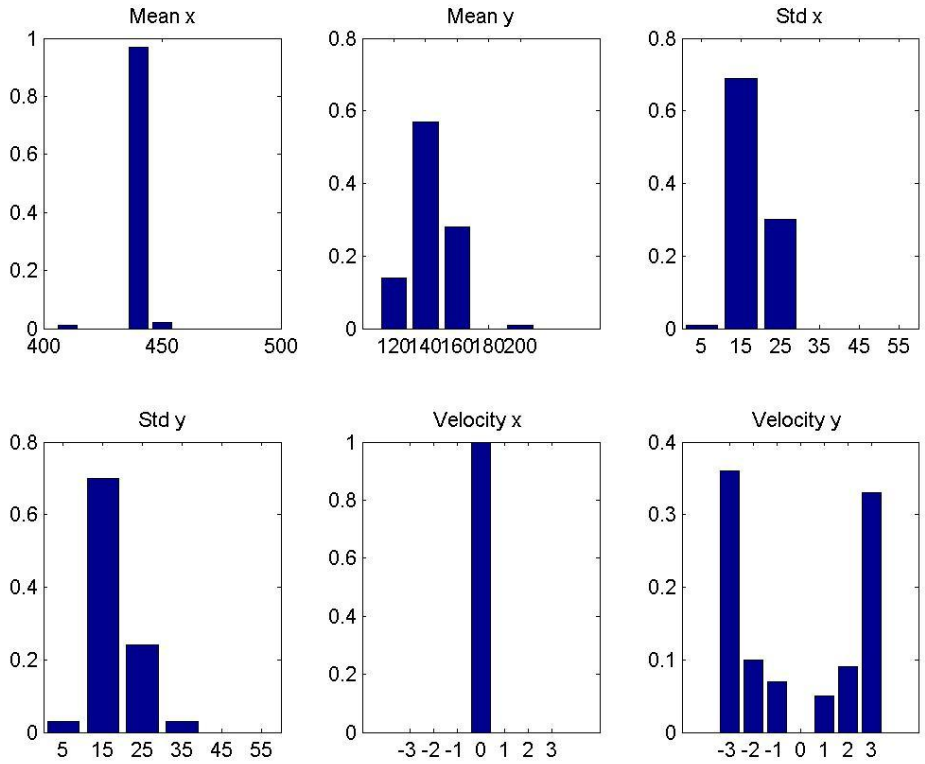

Figure 3.6: Training histogram of stereotype movement *"Body Rocking"*

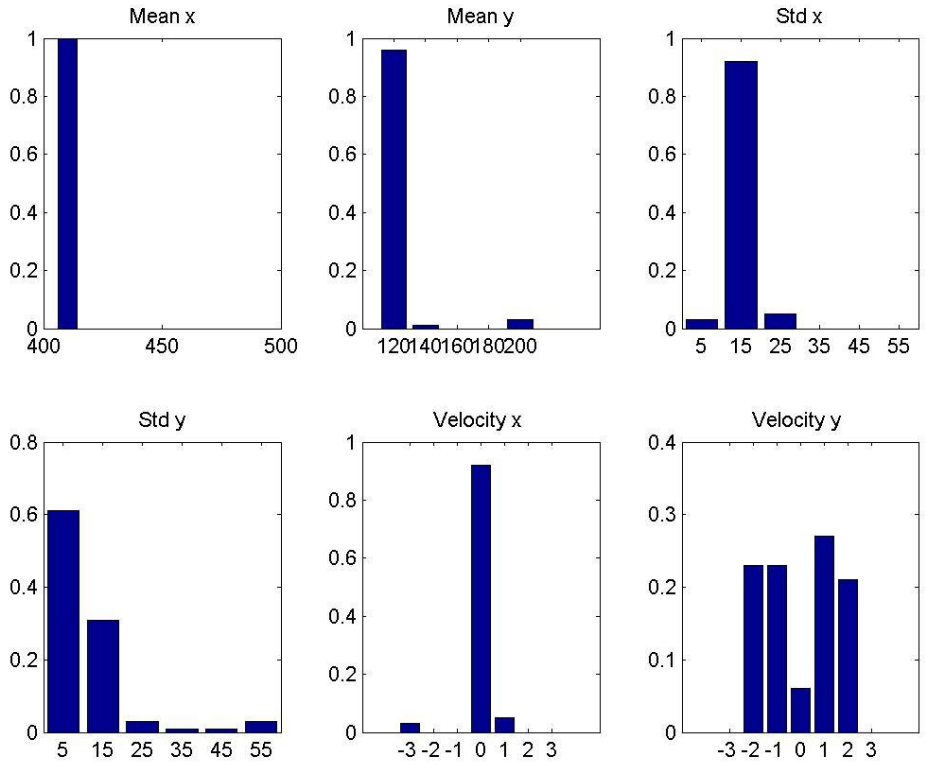

Figure 3.7: Test histogram of stereotype movement *"Nodding"*.

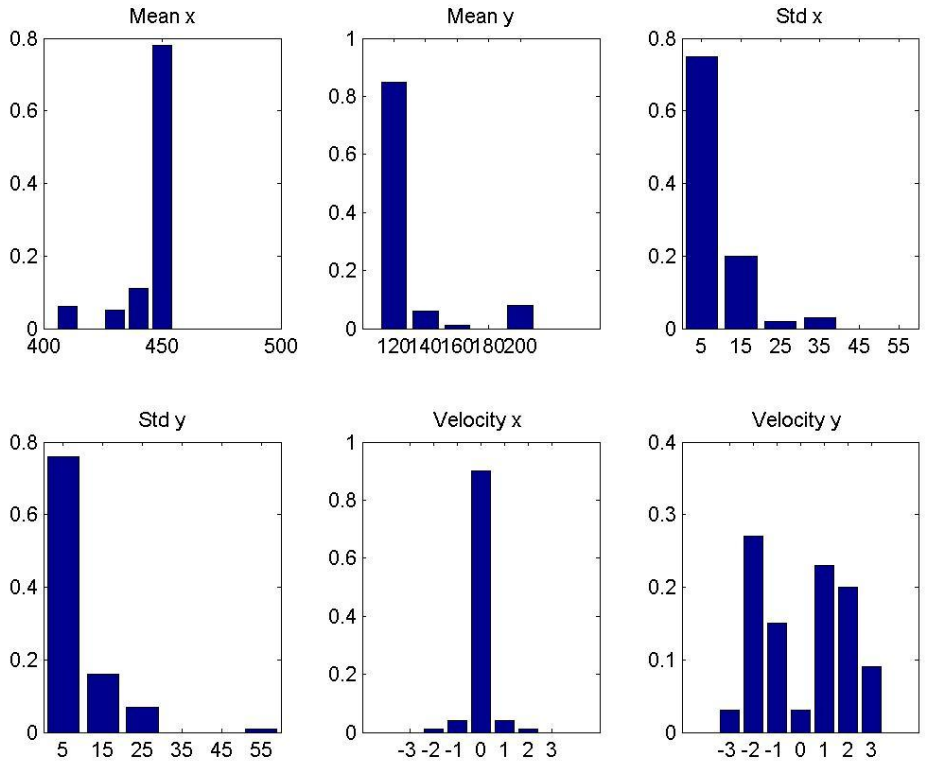

Figure 3.8: Training histogram of stereotype movement *"Nodding"*.

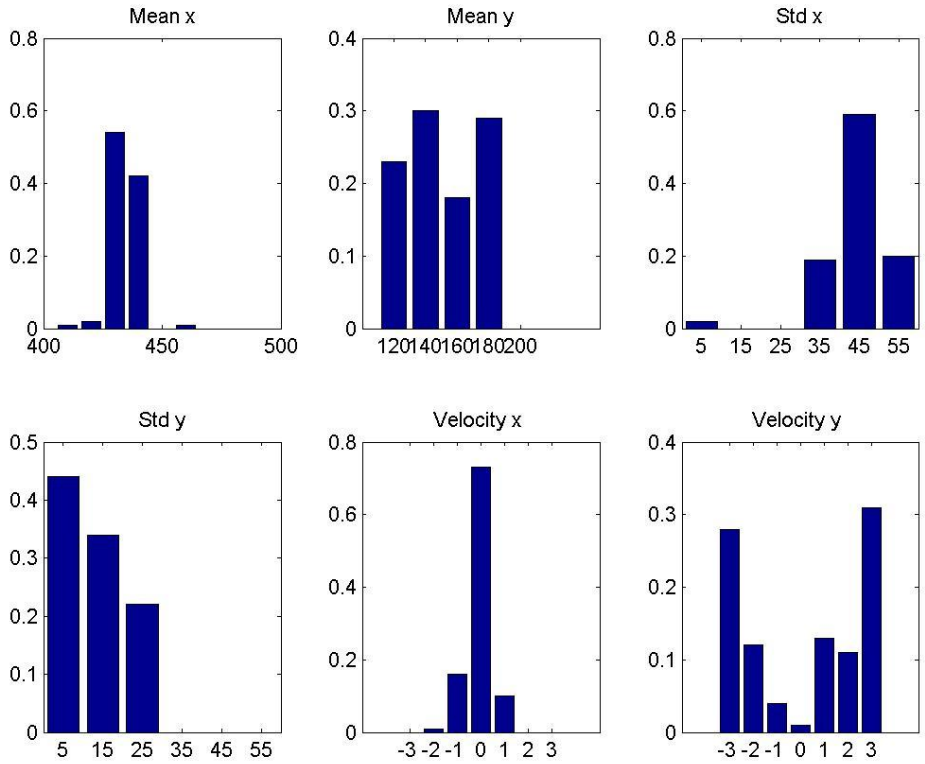

Figure 3.9: Test histogram of stereotype movement *"Hand Flapping"*.

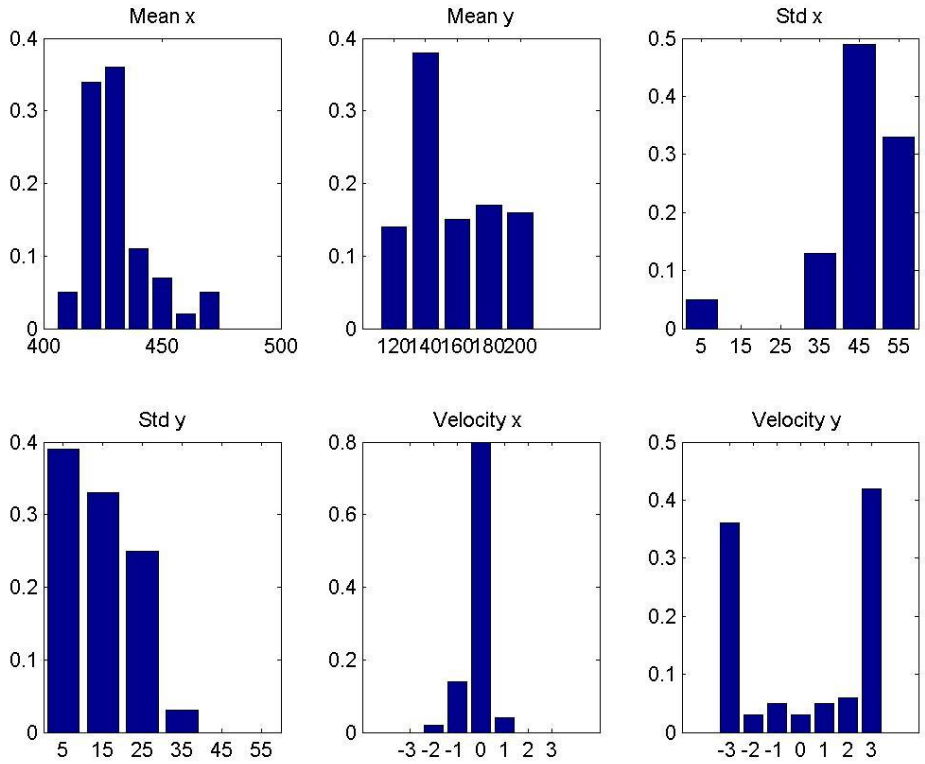

Figure 3.10: Training histogram of stereotype movement *"Hand Flapping"*.

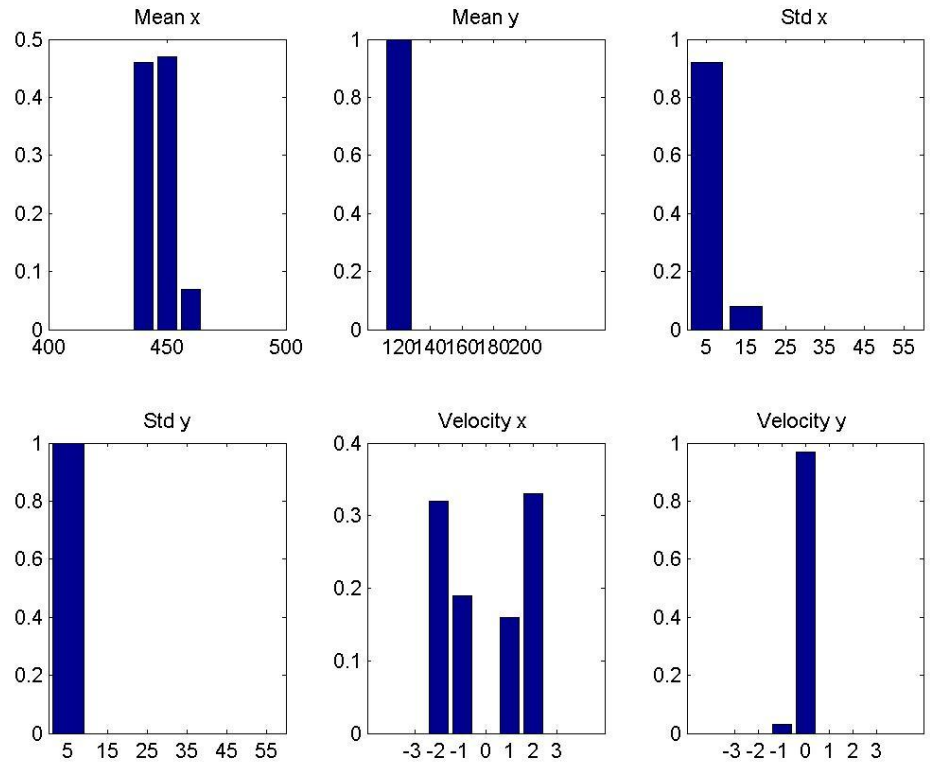

Figure 3.11: Training histogram of stereotype movement *"Shake Head"*.

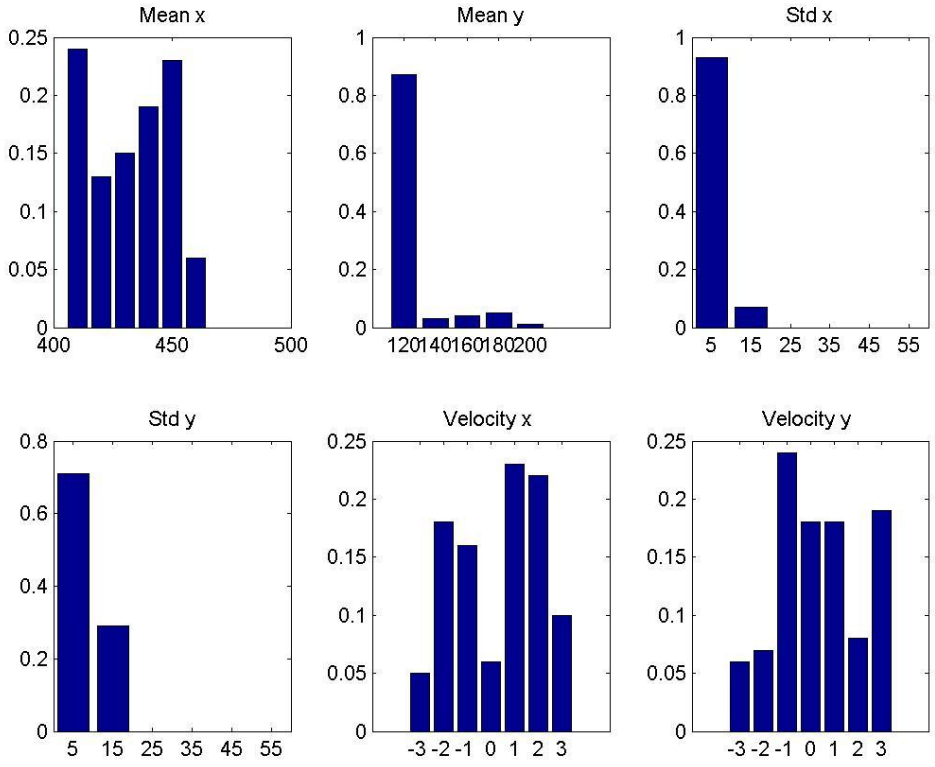

Figure 3.12: Test histogram of stereotype movement *"Hand Wave"*.

3.4 – Analysis of Classification Results

The first tests were conducted using 5 video sequences that were shifted by 10 frames for each binary SVM classifier. Although each one successfully managed to successfully identify their respective movements in their respective frame slot, they had false results during some frames throughout the testing. Movement 3, which corresponds to the stereotype *"Shake Head"*, should not have been classified at all, yet it appeared several times throughout the Testing Sequence. No Movement at all, meaning when all the binary SVM classifiers had an *"Not"* usually appeared between the movement gaps but failed to classify the unknown movement called *"Wave hand"*. In the last sequence no movement was conducted. The following table 3.3 shows the obtained results. The colors were used to enhance the results where green stands for successful classification and red for a classification mistake.

	"Body Rocking"	***"Nodding"***	***"Shake Head"***	***"Hand Flapping"***
"Body Rocking"	*"mov"*	*"not"*	*"mov"*	*"not"*
"Nodding"	*"mov"*	*"mov"*	*"mov"*	*"mov"*
"Wave hand"	*"mov"*	*"mov"*	*"mov"*	*"mov"*
"Hand Flapping"	*"mov"*	*"mov"*	*"mov"*	*"mov"*
"Nothing"	*"mov"*	*"not"*	*"not"*	*"not"*

Table 3.3: Classification results with 5 training samples in the testing sequence.

The percentage of false positives for the first test sequence using 5 training samples, in terms of wrong number of frames per movement section, is represented in Table 3.4.

	"Body Rocking"	*"Nodding"*	*"Shake Head"*	*"Hand Flapping"*	**Sequence Length(Frames)**
"Body Rocking"	0	0	2.25	0	400
"Nodding"	70.45	0	5.67	43.58	335
"Wave hand"	65.09	36.47	3.14	21.96	255
"Hand Flapping"	41.02	21.36	3.58	0	295
"Nothing"	100	0	0	0	137

Table 3.4: False positives with 5 training samples in the testing sequence.

The second test was conducted with 10 training samples that were again shifted by 5 frames for each binary SVM classifier. The following table 3.5 exposes the obtained results.

	"Body Rocking"	***"Nodding"***	***"Shake Head"***	***"Hand Flapping"***
"Body Rocking"	*"mov"*	*"not"*	*"not"*	*"not"*
"Nodding"	*"mov"*	*"mov"*	*"not"*	*"mov"*
"Wave hand"	*"mov"*	*"mov"*	*"mov"*	*"mov"*
"Hand Flapping"	*"mov"*	*"mov"*	*"not"*	*"mov"*
"Nothing"	*"mov"*	*"not"*	*"not"*	*"not"*

Table 3.5: Classification results with 10 training samples in the testing sequence.

The percentage of false positives for 10 training samples is represented in Table 3.6.

	"Body Rocking"	*"Nodding"*	*"Shake Head"*	*"Hand Flapping"*	**Sequence Length(Frames)**
"Body Rocking"	0	0	0	0	400
"Nodding"	31.34	0	0	31.34	335
"Wave hand"	1.18	25.66	38.04	0.39	255

"Hand Flapping"	38.98	18.64	0	0	295
"Nothing"	100	0	0	0	137

Table 3.6: False positives results with 10 training samples in the testing sequence.

Consulting Tables 3.5 and 3.6 allows the conclusion that the increase in training samples helped each SVM classifier to better classify each stereotyped behavior. The overall number of false positives in the classification process has decreased due to the increase in training samples.

The third test was conducted with 20 training samples that were once more shifted by 2 and 3 frames for each binary SVM classifier. The following table 3.7 can demonstrate the classification results.

	"Body Rocking"	***"Nodding"***	***"Shake Head"***	***"Hand Flapping"***
"Body Rocking"	*"mov"*	*"not"*	*"not"*	*"not"*
"Nodding"	*"mov"*	*"mov"*	*"not"*	*"mov"*
"Wave hand"	*"mov"*	*"mov"*	*"mov"*	*"mov"*
"Hand Flapping"	*"mov"*	*"mov"*	*"not"*	*"mov"*
"Nothing"	*"mov"*	*"not"*	*"not"*	*"not"*

Table 3.7: Classification results with 20 training samples in the testing sequence.

The percentage of false positives for 20 training samples is represented in Table 3.8

	"Body Rocking"	***"Nodding"***	***"Shake Head"***	***"Hand Flapping"***	**Sequence Length(Frames)**
"Body Rocking"	0	0	0	0	400
"Nodding"	53.13	0	0	19.40	335

"Wave hand"	25.88	17.65	34.51	0.39	255
"Hand Flapping"	37.29	25.76	0	0	295
"Nothing"	100	0	0	0	137

Table 3.8: False positives results with 20 training samples in the testing sequence.

Tables 3.7 and 3.8 demonstrate again that increase in training samples helped each SVM classifier to improve its classification strength. The overall number of false positives in the classification process has decreased due to the increase in training samples although the stereotyped movement *"Body Rocking"* has increased. This can be due to the frame shifting of the training samples creating empty clusters.

Chapter 4

Conclusions and Future Work

The objectives to achieve were clearly defined at the beginning of this project. The aim was to develop a methodology that would allow the detection of typical stereotype movements of ASD. Analyzing the state of the art in this field allowed to identify certain negative factors that influenced some of the computer vision

approaches. After a careful deliberation it was then decided which techniques and procedures to apply. This led to the development of numerous functional tests and new decisions were made throughout the development of this project.

The approach used in this work was the result of three phases, the first devoted to detect the person's contours using HOG algorithm which later would allow the tracking to be as easier as possible.

The second objective would be to implement a tracking method which allows the person's feature points to be followed in every frame of a given video. The KLT feature point tracker has proven to be very effective and robust in tracking people's features through image sequences and video files.

The last objective of this thesis was to detect routine breaches in autism patients. This would allow them to be assisted immediately through surveillance cameras while not being attended and guarded by medical staff. A database file was filmed specially for this thesis so that the behavior detection could be performed using multiple binary Support vector Machine (SVM) Classification.

The experimental results show that successful classification was obtained even with the limited training examples provided. This leads to the conclusion that the accuracy of the binary SVM classification is directly related to the number of training examples, i.e. the more input data, the more accurate the classification output becomes.

This work, in addition to scientific interest and medical purpose, has practical utility. It can help children at an early stage in their lives to be treated for ASD and allow them to have a more dignifying life.

A possible improvement of this project would be the implementation in real-time. The migration to another more versatile programming language like C or C++ would meet the requirements that a real-time application requires.

In case of a bigger number of stereotypes, using the table described in appendix A.1 for example, LibSVM could be used instead of MatLab's SVM since it allows native multiclass classification.

Another possible improvement to this work could be the use of the Maximization of Mutual Information (MMI) algorithm.

The combination of the KLT tracking algorithm with Extended Kalman Filter allows to minimize tracking errors and could prove to be a valuable asset in feature tracking approaches [77].

As stated in the State of the Art, the two mentioned types of object classification have proven to obtain more robust classifiers [78]. Hybrid Classifiers have shown a significant superiority compared to those purely based on appearance or on separate movements [79, 80]. The fusion of multiple features is becoming very promising approach for classification of objects in realistic scenarios.

The Microsoft Kinect Sensor (MKS) seems to be an interesting choice equipment wise. Its depth sensor has the ability to identify the major joints of the human body in a non-intrusive and precise way. The MKS equipment also allows good software development conditions for a relatively small price.

References

[1] Sudarsun. S, Varun Kant Vashishtha, and Avijit Nayak *"Using Behavioral Patterns in Treating the Autistic"*, 2007.

[2] Mohamed-Bécha Kaâniche,François Brémond ,"*Recognizing Gestures by Learning Local Motion Signatures of HOG Descriptors"*,February 22, 2012.

[3] James M. Rehg*,"Behavior Imaging:Using Computer Vision to Study Autism".*In Conference on Machine Vision Applications,June 13-15-2011.

[4] J. K. Aggarwal, J. Park, S. Park, *"Model based Human Motion Tracking and Behavior Recognition using Hierarchical Finite State Automata"*,in ICCSA 2004,page 311-320.

[5] W. Hu, T. Tan, L. Wang, *"Recent development in human motion analysis"*, in Pattern Recognition, page 585-601, 2003.

[6] Robert P.W. Duin, J. Novovičová, P. Pavlík, *"A trainable similarity Measure for Image Classification"*, IEEE, 2006.

[7] J. Hashemi, J. Thiago, V. Spina, M. Tepper, A. Esler, V. Morellas, N. Papanikolopoulos and G. Sapiro, *"Computer vision tools for the non-invasive assessment of autism-related behavioral markers",* 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics, ICDL-EPIROB 2012, San Diego, CA, USA, November 7-9, 2012.

[8] R. Gross and J. Shi, *"The CMU motion of mody (MoBo) database"*, no. CMU-RI-TR-01-18, June 2001.

[9] 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, MARCH 2001.

[10] J. Rittscher, A. Blake, and S. Roberts, *"Towards the automatic analysis of complex human body motions"*, Image and Vision Computing, vol. 20, no. 12, pp. 905–916, 2002.

[11] H. Yu, G. Sun, W. Song, and X. Li, *"Human motion recognition based on neural network"*,Proc. IEEE Conf. Communications, Circuits and Systems, vol. 2, pp. 977–982, 2005.

[12] H. Chen, H. Chen, Y. Chen, and S. Lee, *"Human action recognition using star skeleton"*,Proc. the 4th ACM international workshop on Video surveillance and sensor networks, pp. 171–178, 2006.

[13] H. Li, S. Lin, Y. Zhang, and K. Tao, *"Automatic video-based analysis of athlete action"*, Proc. IEEE Conf. Image Analysis and Processing, pp. 205–210, 2007.

[14] O. Masoud and N. Papanikolopoulos, *"A method for human action recognition",* Image and Vision Computing, vol. 21, no. 8, 2003.

[15] M. Yong, K. Yoshinori, K. Koichi et al., *"Sparse Bayesian Regression for Head Pose Estimation"*,Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, pp. 507-510, 2006.

[16] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard, *"Tracking loose-limbed people"* ,Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 421–428, 2004.

[17] F. Caillette, A. Galata, and T. Howard, *"Real-time 3-D human body tracking using variable length markov models"*, Proc.British Machine Vision Conference, pp. 469–478, 2005.

[18] C. Menier, E. Boyer, and B. Raffin, *"3D skeleton-based body pose recovery"*,Proc. Int. Symp. 3D Data Processing, Visualization and Transmission, pp. 389–396, 2006.

[19] A. Fossati, M. Dimitrijevic, V. Lepetit, and P. Fua, *"Bridging the gap between detection and tracking for 3D monocular video-based motion capture"*,Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8, 2007.

[20] D. Weinland, R. Ronfard, and E. Boyer, *"Free viewpoint action recognition using motion history volumes"*, Computer Vision and Image Understanding, vol. 104, pp. 249–257, 2006.

[21] A. Yilmaz and M. Shah, *"Actions sketch: a novel action representation"*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 984–989, 2005.

[22] F. Lv and R. Nevatia, *"Single view human action recognition using key pose matching and viterbi path searching"*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8, 2007.

[23]. D. Weinland, F. Grenoble, E. Boyer, R. Ronfard, and A. Inc, *"Action recognition from arbitrary views using 3D exemplars"*, Proc. IEEE Conf. Computer Vision, pp. 1–7, 2007.

[24] W. Hu, T. Tan, L. Wang, and S. Maybank, *"A survey on visual surveillance of object motion and behaviors"*, IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 34, no. 3, pp. 334–352, 2004.

[25] P. Viola, M. Jones, *"Robust Real-time Object Detection"*, International Journal of Computer Vision, vol. 2, pp. 882–888, 2001.

[26] A. Lipton, H. Fujiyoshi, and R. Patil, *"Moving target classification and tracking from real-time video"*, Proc. IEEE Workshop. Applications of Computer Vision, pp. 8–14, 1998.

[27] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 4, pp. 349–361, 2001.

[28] Q. Zhou and J. Aggarwal, *"Tracking and classifying moving objects from video",* Proc. IEEE Workshop on Performance Evaluation of Tracking and Surveillance, pp. 1–8, 2001.

[29] O. Javed and M. Shah,*"Tracking and object classification for automated surveillance"*, Proc. European Conf. Computer Vision-Part IV, pp. 343–357, 2002.

[30] E. Rivlin, M. Rudzsky, R. Goldenberg, U. Bogomolov, and S. Lepchev, *"A real-time system for classification of moving objects"*, Proc. IEEE Conf. Pattern Recognition, vol. 3, pp. 688–691, 2002.

[31] M. Rodriguez and M. Shah, *"Detecting and segmenting humans in crowded scenes"*, Proc.Int.Conf. Multimedia, pp. 353–356, 2007.

[32] R. Cutler and L. Davis, *"Robust real-time periodic motion detection, analysis, and application"*, IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 781–796, 2000.

[33] J. Aggarwal and Q. Cai, *"Human motion analysis: A review"*, Computer Vision and Image Understanding, vol. 73, no. 3, pp. 428–440, 1999.

[34] 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 MARCH 2001.

[35] A. Liu, A. Pentland, *"Modeling and Prediction of Humor Behavior*,Neural Computation 11, pp. 229-242, 1999.

[36] P. Patrick and W. Svetha, Vand Geoff, *"Tracking as recognition for articulated full body human motion anlysis",* Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8, 2007.

[37] A. Elgammal, V. Shet, Y. Yacoob, and L. Davis, *"Learning dynamics for exemplar-based gesture recognition"* , Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 571–578, 2003.

[38] T. Mori, Y. Segawa, M. Shimosaka, and T. Sato, *"Hierarchical recognition*" , 2004.

[39] Y. Shi, A. Bobick, and I. Essa, *"Learning temporal sequence model from partially labeled data"*, Computer Vision and Pattern Recognition, pp. 1631–1638, 2006.

[40] F. Lv and R. Nevatia, *"Recognition and segmentation of 3-D human action using HMM and multi-class AdaBoost"* , Proc. European Conf. on Computer Vision, vol. 4, pp. 359–372, 2006.

[41] A. Fossati, M. Dimitrijevic, V. Lepetit, and P. Fua, *"Bridging the gap between detection and tracking for 3D monocular video-based motion capture"* , Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8, 2007.

[42] C. Curio and M. Giese, *"Combining view-based and model-based tracking of articulated human movements"*, Proc. IEEE Workshop on Motion and Video Computing, vol. 2, pp. 261–268, 2005.

[43]] R. Natarajan, P. Nevatia, *"View and scale invariant action recognition using multiview shape-flow model"* , Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8, 2008.

[44] Y. Ming-Hsuan, D. J. Kriegman, and N. Ahuja, *"Detecting Faces in Images: A Survey"* , Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 1, pp. 34-58, 2002.

[45] G. Yang, and T. S. Huang, *"Human face detection in a complex background"*, Pattern Recognition, vol. 27, no. 1, pp. 53-63, 1994

[46] M.Yang, and N. Ahuja, *"Face Detection and Gesture Recognition for Human-Computer Interaction (The International Series in Video Computing)"*, pp 1-52, Springer, 2001.

[47] Y. Dai, and Y. Nakano, *"Face-texture model based on SGLD and its application in face detection in a color scene",* Pattern Recognition, vol. 29, no. 6, pp. 1007-1017, 1996.

[48] F. Bayoumi, M. Fouad, and S. Shaheen, *"Feature-based human face detection*", Radio Science Conference, 2004. NRSC 2004. Proceedings of the Twenty-First National, pp. C21-1-10, 2004.

[49] I. Craw, D. Tock, and A. Bennett, *"Finding Face Features"* , in Proceedings of the Second European Conference on Computer Vision, pp. 92-96, 1992.

[50] A. Lanitis, C. J. Taylor, and T. F. Cootes, *"Automatic face identification system using flexible appearance models"*, Image and Vision Computing, vol. 13, no. 5, pp. 393-401, 1995.

[51] M. Turk, and A. Pentland, *"Eigenfaces for recognition"*, J. Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991.

[52] H. Rowley, S. Baluja, and T. Kanade, "*Neural Network-Based Face Detection",* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38, 1998.

[53] A. N. Rajagopalan, K. S. Kumar, J. Karlekar, *"Finding faces in photograph"*, Computer Vision, 1998. Sixth International Conference on, pp. 640-645, 1998.

[54] S. Niyogi, and W. T. Freeman, *"Example-based head tracking"* , Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on, pp. 374-378, 1996.

[55] D. J. Beymer, *"Face recognition under varying pose"*, Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94, 1994 IEEE Computer Society Conference on, pp. 756-761, 1994.

[56] J. Huang, X. Shao, and H. Wechsler, *"Face Pose Discrimination Using Support Vector Machines (SVM)"*, Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on, pp. 154-156 vol.1, 1998.

[57] M. Jones, and P. Viola, *"Fast multi-view face detection"*, Mitsubishi Electric Research Laboratories, 2003.

[58] Y. Li, S. Gong, J. Sherrah, *"Support vector machine based multi-view face detection and recognition"*, Image and Vision Computing, vol. 22, no. 5, pp. 413-427, 2004.

[59] M. Yong, K. Yoshinori, K. Koichi et al., *"Sparse Bayesian Regression for Head Pose Estimation"*, Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, pp. 507-510, 2006.

[60] M. Hankyu, and M. L. Miller, *"Estimating facial pose from a sparse representation [face recognition applications]"*, Image Processing, 2004. ICIP '04. 2004 International Conference on, pp. 75-78 Vol. 1, 2004.

[61] J. B. Tenenbaum, V. de Silva, and J. C. Langford, *"A Global Geometric Framework for Nonlinear Dimensionality Reduction"*, American Association for the Advancement of Science, vol. 290, no. 5500, pp. 2319-2323 2000.

[62] V. N. Balasubramanian, Y. Jieping, and S. Panchanathan, "Biased Manifold Embedding: A Framework for Person-Independent Head Pose Estimation," Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pp. 1-7, 2007.

[63] J. Sherrah, S. Gong, and E. J. Ong, *"Face distributions in similarity space under varying head pose",* Image and Vision Computing, vol. 19, no. 12, pp. 807-819, 2001.

[64] S. Srinivasan, and K. L. Boyer, "Head pose estimation using view based eigenspaces," Pattern Recognition, 2002. Proceedings. 16th International Conference on, pp. 302-305 vol.4, 2002.

[65] M. Belkin, and P. Niyogi, "*Laplacian Eigenmaps for Dimensionality Reduction and Data Representation"*, Neural Computation, vol. 15, no. 6, pp. 1373-1396, 2003.

[66] Z. Biuk, and S. Loncaric, *"Face recognition from multi-pose image sequence"*, Image and Signal Processing and Analysis, 2001. ISPA 2001. Proceedings of the 2nd International Symposium on, pp. 319-324, 2001

[67] T. F. Cootes, G. V. Wheeler, K. N. Walker et al., "*View-based active appearance models"*, Image and Vision Computing, vol. 20, no. 9–10, pp. 657-664, 2002.

[68] X. Jing, S. Baker, I. Matthews et al., *"Real-time combined 2D+3D active appearance models"*, Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, pp. II-535-II-542 Vol.2, 2004.

[69] A. Gee, and R. Cipolla, *"Determining the gaze of faces in images"*, Image and Vision Computing, vol. 12, no. 10, pp. 639-647, 1994.

[70] T. Horprasert, Y. Yacoob, and L. S. Davis, *"Computing 3-D head orientation from a monocular image sequence"*, Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on, pp. 242-247, 1996.

[71] M. E. Timmerman, *"Principal Component Analysis (2nd Ed.). I. T. Jolliffe"*, Journal of the American Statistical Association, vol. 98, pp. 1082-1083, 2003.

[72] Shi, J. and Tomasi, C.,*"Good Features to Track"*, IEEE Conference on Computer Vision and Pattern Recognition, 1994.

[73] Dalal, N. and B. Triggs. *"Histograms of Oriented Gradients for Human Detection"*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition,Vol. 1, pp. 886–893, June 2005.

[74] Golparvar-Fard, M., Arsalan Heydarian, A., Niebles, J. C., *"Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers",* Advanced Engineering Informatics 27,pp. 652–663, 2013.

[75] V. Vapnik, L. Bottou, *"On structural risk minimization or overall risk in a problem of pattern recognition"*, Automation and Remote Control, 1977.

[76] L. Yann, J.L.D.E. Harris, B.N.C. Corinna, D.J.S.D. Harris, S. Eduard, S. Patrice, V. Vladimir, *"Learning algorithms for classification: a comparison on handwritten digit recognition neural networks"*,The Statistical Mechanics Perspective, pp. 261–276, 1995.

[77] Kâaniche, M. D., "*Human Gesture Recognition"*, Ph.D. Thesis, October 2009.

[78] L. Oliveira, U. Nunes and P. Peixoto, *"On Exploration of Classifier Ensemble Synergism in Pedestrian Detection"*, IEEE Transactions on Intelligent Transportation Systems, Vol. 11, N. 1, March 2010.

[79] Y. Bogomolov, G. Dror, S. Lapchev, E. Rivlin, M. Rudzsky, and I. Tel-Aviv, *"Classification of moving targets based on motion and appearance"*, Proc. British Machine Vision Conference, vol. 2, pp. 429–438, 2003.

[80] P. Viola, M. Jones, and D. Snow, *"Detecting Pedestrians using Patterns of Motion and Appearance"*, 2005.

[81] J. B. MacQueen ,"*Some Methods for classification and Analysis of Multivariate Observations"*,in Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297, 1967.

[82] C. Yildiz,*"An Implementation on Histogram of oriented Gradients for Human Detection"*,20??.

[83] Lam, K. S. L. and M. G. Aman, "*The Repetitive Behavior Scale-Revised: independent validation in individuals with autism spectrum disorders*", Journal of Autism and Developmental Disorders, 37(5): 855-866, 2007.

[84] Rosten, E., and T. Drummond. *"Fusing Points and Lines for High Performance Tracking," Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2, pp. 1508–1511, October 2005.

[85] Harris, C., and M. Stephens, *"A Combined Corner and Edge Detector" Proceedings of the 4th Alvey Vision Conference*, pp. 147-151, August 1988.

[86] Bay, Herbert, Andreas Ess, Tinne Tuytelaars, Luc Van Gool, SURF: *"Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU)*, Vol. 110, No. 3, pp. 346--359, 2008.

[87] Nister, D., and H. Stewenius, *"Linear Time Maximally Stable Extremal Regions", Lecture Notes in Computer Science*. 10th European Conference on Computer Vision, Marseille, France: 2008, no. 5303, pp. 183–196.

[88] Seber, G. A. F. "*Multivariate Observations"*, Hoboken, NJ: John Wiley & Sons, Inc., 1984.

Appendix A

Stereotype Identification

The ASD demonstrates the particularity of showing repetitive and restrictive behaviors besides limited social interaction and communication. These behaviors often assume a stereotypic motor movement which, besides having a non-existent definition, corresponds to frequent motor movements described as rhythmic, pointless and involuntary.

Restrictive stereotype movements interfere with the patient's autonomy and their ability to perform other activities. Stereotypes are also inconvenient for social integration since they are seen as inadequate and uncommon.

Stereotype classification is frequently based on the distinctions used in the RBS-R [83]. Some of them are described in table A.1.

Body Part	**Stereotype Movement Description**
Whole Body	Body rocking
	Body swaying
Head	Rolls head
	Nods head
	Turns head
Hand or Finger	Flaps hands
	Wiggles or flicks fingers
	Claps hands
	Waves or shakes hand or arms
Locomotion	Turns in circles

	Whirls
	Jumps
	Bounces
Object usage	Spins or twirls objects
	Twiddles or slaps or throws objects
	Let’s objects fall out of hands
Sensory	Covers eyes
	Looks closely or gazes at hands or objects
	Covers ears
	Smells or sniffs items
	Rubs surfaces

Table A.1: Stereotype movement list from RBS-R.