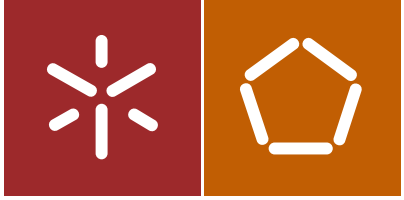
Universidade do Minho
Escola de Engenharia

Patrícia Margarida Silva de Castro Neves Barbosa

Human features detection in video surveillance

October 2016

UMinho | 2016 Patrícia Margarida Silva de Castro Neves Barbosa Human features detection in video surveillance



Universidade do Minho
Escola de Engenharia

Patrícia Margarida Silva de Castro Neves Barbosa

Human features detection in video surveillance

Master's dissertation in
Engineering Industrial Electronics and Computers

Work performed under the guidance of/from/to
Professor Filomena Maria da Rocha Menezes de Oliveira Soares
Professor Henrique Manuel Dinis dos Santos

October 2016

ACKNOWLEDGMENTS

The first words of thanks are directed to my parents, Elísio Barbosa e Margarida Barbosa, during the educational, psychological and financial support provided through all my academic development, because, without them, it will be impossible to accomplish this huge step in my life.

Special thanks to my brother, Miguel Barbosa, and my sister, Liliana Barbosa, for their tireless encouragement, friendship and disposition to help me, in the good and bad times.

I thank my advisor's professor Filomena Soares, professor Henrique Santos and engineer Duarte Duque for all their enthusiasm and support. Without them, I wouldn't know the best way to accomplish the various stages of this dissertation.

I thank EXVA technologies for the opportunity to develop my dissertation for their company.

I also want to thank, a good friend of mine, Sara Costa, for all the moments of friendship lived in this past six years, without her sincerity and kindness it will be much more difficult to overcome this period in my life.

Finally, but not least, my boyfriend, Rafael Martins, for giving me joyful moments and for the mainly support in moments of anguish and disappointment.

To all, thank you!

ABSTRACT

Human activity recognition algorithms have been studied actively from decades using a sequence of 2D and 3D images from a video surveillance. This new surveillance solutions and the areas of image processing and analysis have been receiving special attention and interest from the scientific community. Thus, it became possible to witness the appearance of new video compression techniques, the transmission of audio and video in real-time, targeting identification and tracking objects in with complex environments. Traffic monitoring, automotive safety, people counting and activity recognition applications are examples. With the development of sensors, new opportunities arose to expand and advance this field.

This dissertation presents an activity recognition system to recognize human motion. The system does not need optical markers or motion sensors. This human activity recognition system is divided in three stages: human segmentation, in an outside and inside environment; extraction of the human features; and classification models to detect the human actions. Therefore, the main objective in this work is to develop an algorithm to extract human features. This algorithm aims to develop a new representation and extraction method using a sequence of features in a skeleton silhouette. Mainly, the segmentation of humans is based on a previous work, centered on background subtraction. An algorithm is applied to convert the object captured in the video surveillance to a binary image using a skeleton algorithm. Afterwards, and based on the physical parameters of the human motion, it becomes possible to discover the principal features of the human skeleton, called physical features, head, hands and feet.

The viability of using features detection in a human recognition system was tested and compared with other existing systems. The results point out that the system has good performance (8.96% of perfect match and the average rate was 68.65%). Nevertheless, in images where the features of the human body are covered, with umbrella or heavy coats for example, the system presents certain limitations. This process has a high execution speed and a low cost computational processing: average of 5910 μs with a standard deviation of 5650 μs.

In the near future, classification models to detect the human actions will be included.

KEYWORDS: VIDEO SURVEILLANCE, HUMAN FEATURES SELECTION, SKELETON

RESUMO

Algoritmos de reconhecimento de atividade humana foram estudados ativamente durante décadas, usando sequências de imagens em 2D e 3D de vídeo vigilância. Estas novas soluções de vídeo vigilância e as áreas de processamento e análise de imagens têm recebido especial atenção e interesse por parte da comunidade científica. Assim, tornou-se possível testemunhar a aparência de novas técnicas de compressão de vídeo, a transmissão de áudio e vídeo em tempo real, identificação de segmentação e rastreamento de objetos em ambientes complexos. Monitoramento de tráfego, segurança automóvel, contagem de pessoas e aplicações de reconhecimento de atividade são exemplos. Com o desenvolvimento de sensores, novas oportunidades surgiram para expandir e avançar neste campo.

Esta dissertação apresenta um sistema de reconhecimento de atividade para reconhecer o movimento humano. O sistema não precisa de marcadores óticos ou sensores de movimento. Este sistema de reconhecimento de atividade humana divide-se em três fases: segmentação humana, num ambiente exterior e interior; Extração das características humanas; E modelos de classificação para detetar as ações humanas. Portanto, o objetivo principal deste trabalho trata-se de desenvolver um algoritmo para extrair características humanas. Este algoritmo tem como objetivo desenvolver uma nova representação e método de extração de características humanas, através do uso de uma silhueta em forma de esqueleto. A segmentação de seres humanos é baseada num trabalho anterior, centrado na subtração do plano de fundo. Um algoritmo é aplicado para converter o objeto capturado na vídeo vigilância, para uma imagem binária usando um algoritmo em forma de esqueleto. Posteriormente, e com base nos parâmetros físicos do movimento humano, torna-se possível descobrir as principais características do esqueleto humano, denominadas características físicas, cabeça, mãos e pés.

A viabilidade do uso de deteção de características em um sistema de reconhecimento humano foi testada e comparada com outros sistemas. Os resultados indicam que o sistema tem bom desempenho (8.96% de correspondência exata e 68.65% de correspondência intermédia). No entanto, em imagens onde as características do corpo humano são cobertas, com guarda-chuva ou casacos pesados, por exemplo, o sistema apresenta certas limitações. Este processo tem uma alta velocidade de execução e um processamento computacional de baixo custo: média de 5910 μs com desvio padrão de 5650 μs.

Num futuro próximo, serão incluídos modelos de classificação para detetar as ações humanas.

PALAVRAS-CHAVE: VÍDEO VIGILÂNCIA, CARACTERÍSTICAS HUMANAS, ESQUELETO

Contents

LIST OF FIGURES

LIST OF TABLES

List of Acronyms

2D – Dimensional

3D – Tridimensional

BN - Bayesian Network

BBN - Bayesian Belief Network

BOF - Bag of features

CMOS – Complementary Metal-Oxide Semiconductor

DT - Temporal Differencing

FSF - Free Software Foundation

GNU – Gnu Not Unix

HCI - Human-computer interaction

HMM - Hidden Markov model

HSL – Hue, Saturation, Lightness

HSV – Hue, Saturation, Value

KDE - Kernel density estimation

MAP – Maximum a Posteriori

MATLAB – MATrix LABoratory

MLE - Maximum Likelihood Estimation

MOG – Mixture of Gaussians

OCR - Optical character recognition

RGB – Red-Green-Blue

RGB-D – Red-Green-Blue-Depth

ROI - Region of interest

SMIJ – Sequence of Most Informative Joints

SVM - Support Vector Machine

W^4 – Who? When? Where? What?

1. INTRODUCTION

This chapter aims to present the context in which this work is inserted, and the reasons that led to the development of this dissertation. Also during the introductory chapter, the main contributions will be identified, as well as the objectives and the methodology adopted for this study.

1.1 Context

The security in public and private spaces is intended to detect and prevent unauthorized events. Usually, this security is mostly performed by security officers in the field, but in certain situations, the relationship between cost and benefit reaches large values. In these cases, the use of dynamic video surveillance equipment systems allows human resources retrenchment, through a remote and surveillance. Therefore, a computer vision based solution would make it possible to monitor some, if not all, video sources at the same time to track events that require further human interaction, to alert the user if an unusual event occurs so he can take further action. For example, the system could instantly identify all scene participants, track some suspects and quickly consider their previous activities.

Forsyth et al. [1] designated computer vision as an understanding of cameras and the physical process of image formation to obtain simple inferences from individual pixel values, combining the information available in multiple images into a coherent result, imposing some order on groups of pixels to separate them from each other or infer shape information, and recognize objects using geometric information or probabilistic techniques.

In the past few years there has been a significant positive evolution on processing capacity. Because of this technological evolution, capture techniques, processing and scanned images analysis also noticed a huge improvement.

Nowadays, action recognition has become a very important subject in computer vision, with some vast applications, such as, video surveillance, multimedia, human-computer interfaces and robotics, among others [2]. However, detecting and analyzing action recognition in real time from video algorithms became viable only recently: *Pfinder*[3] and W^4 algorithm [4] are examples. However, they could not solve the problem of recognizing and analyzing humans, because they encountered some drawbacks that it will be discussed in section 2.5 [5].

Feature extraction consists in detecting human motion and posture and use them to analyze human action types, with different kind of representations, from complex body models to just a silhouette image [2]. Therefore, visual observations, from video cameras, has the greatest part in this thesis,

however it should be noted that actions can usually be predicted from other sensory channels, such as audio [2].

After the extraction and collection of the human features, the next step is to classify, using some learning statistical models to determine the actions of the subjects. This kind of actions categories can be difficult to identify in practice, since it must identify for each action the characteristic attitude, such as, running, punching, among others, while preserving appropriate adaptability to all forms of variations. Consequently, vision-based techniques can be classified accordingly to many different criteria, such as the body parts involved (facial expressions, hand gestures, upper-body gestures, full-body gestures, among others); the selected image features (interest points, landmarks, edges, optical flow, and others); and the class of statistical models used for learning and recognition (nearest neighbors, discriminate analysis, among others) [2].

1.2 Main objective of this dissertation

The main objective of this dissertation is the development of a human action recognition system using a surveillance camera. It intends to create a new representation and extraction method using a sequence of features in a skeleton silhouette. Consequently, in this dissertation, algorithms for detecting human features from a sequence of images in a constrain camera were investigated. Also, algorithms for segmentation and classification of human actions were also studied.

Thus, this dissertation has the following tasks:

i. Study the better method to track human motion in a video surveillance;
ii. Investigate the different types of human segmentation, identifying the most appropriate;
iii. Consider different machine learning algorithms;
iv. Study and analyze the different types of algorithms to detect human features in a sequence of images;
v. Investigate other existing systems for human motion recognition;
vi. Develop a system to detect human features;
vii. Discuss the results of the project, by comparing with other projects;

1.3 Organization of the dissertation

This dissertation is structured in four chapters.

The second chapter presents several fundamental aspects indispensable for the comprehension of this work, such as, object segmentation techniques and some algorithms for detecting the human features that have already been developed in the past thirty years. It is also discussed some advantages and disadvantages of each case.

In chapter three it is presented an alternative to resolve tracking and human motion analysis problem, based on detecting features in a skeletonization of the human body, more precisely the head, hands and feet. It is also discussed the benefits of using this technique to resolve the human motion recognition problem.

The results obtained are presented and analyzed in chapter four.

The dissertation concludes with chapter five, which presents the main conclusions of this dissertation, as well as some suggestions for future work.

2. STATE OF ART

This dissertation is framed in the field of human features detection in video surveillance, and therefore this chapter intents to provide an overview on the already existing methods. Firstly, it will be exposed the general concepts on this theme. Secondly, it will be presented the most relevant efforts made by the science community, presenting the best human features detection methods for this case.

2.1 Human action recognition

The challenge of detecting and tracking moving objects has been addressed by several researchers trying to accomplish the best performance on this problem. The detection of human action is a complex problem when dealing with uncontrolled environments, such as outdoor scenes. In such scenarios, changes in light intensity, deformation, shadows and interaction between objects are fairly common issues that normally are a source of noise in the system.

In human action recognition, there are many ways of structuring the different disciplines of human motion capture. However, the most notable technique was introduced by *Moeslund et al.* [6], in 2001. This structure consists of four steps, as we can see in Figure 1. First, it is initialized the data preprocessing, this step covers the actions needed to guarantee that a system begins its operation with a precise interpretation of the current scene. Initialization mainly apprehensions are camera calibration, adaption to scene characteristics, and model initialization, for example, a proper model of the system should be built.

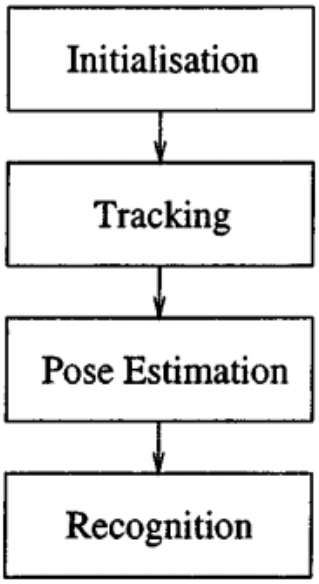


Figure 1- Moeslund et al. structure for systems analyzing human body motion [6]

After initialization, it is necessary to implement segmentation of the object from the background, foreground segmentation, enabling tracking consecutive frames, named the tracking step. In this step, must consider three important aspects. Firstly, nearly every tracking algorithm within human motion capture starts with the same problem: segmenting the human figure from the rest of the image. Secondly, these segmented images are transformed into another representation to reduce the amount of information or to suit a particularly algorithm. Third, the frame to frame object tracking method has to be defined [6].

After the tracking is concluded, it is important to identify how a human body or individual limbs are configured in a certain scene, pose estimation, which could be used in a surveillance system or in a human-computer interaction system (HCI). The human model created in this step, can be used to estimate the precise pose in terms of positions, orientation, width, velocity, etc.

The final step is called recognition. It is usually used to classify the captured motion as one of several types of actions. However, the actions are usually simple, like walking and running, there has been an enormous advance in detecting difficult actions, such as ballet dance steps for example.

As mentioned above, this structure presents the steps for creating a system for human action recognition, despite that, not all the systems require all steps, *Moeslund et al*. guaranteed that more than 130 human motion capture papers published since 1980 to 2002 were reviewed based on their work [6].

This works pretends to develop an algorithm for "Pose Estimation". Therefore, to understand the present work only different kinds of "tracking" and "pose estimation" methods were addressed.

2.2 Object segmentation approaches

The aim of segmentation is to identify the semantically meaningful components of an image and grouping the pixels belonging to such components. It is more practical to segment moving objects from dynamic scene with the aid of motion information contained in it [7]. Segmentation of moving objects in image sequence plays an important role in image sequence processing and analysis. Once the moving objects are detected or extracted out, they can serve for varieties of purposes [7]. Therefore, in this section it were presented some types of approaches to discriminate moving objects on the video stream, the second step in *Moeslund et al*. technique [6].

Nowadays, the most common methods implemented to detect moving targets are temporal differencing [8], optical flow [9] [10] [11] and background subtraction [4].

2.2.1 Temporal differencing

A basic method for target tracking in real-time video applications is temporal differencing (DT). This method calculates sequential images, separated by constant frames of time, to find regions which have undergone a change. The processed images may differ in color space, which may be considered grayscale images or color, for example, RGB, HSV and HSL.

Temporal differencing uses a point to designated as belonging to a region of foreground. However, if the absolute difference of the value of that point between two sequence images is greater than a predetermined an empirical value, threshold. This value can be determined by using a threshold function, depicted in the equation in (eq. 1). This equation shows that if I_n is the intensity of the n^{th} frame then the pixel wise difference function Δ_n [12].

$$\Delta_n = |I_n - I_{n-1}| \quad (1)$$

Then, to determinate the region of motion in an image it has to use the (eq. 2)[12]:

$$M_n(u,v) = \begin{cases} I_n(u,v) & , \quad \Delta_n(u,v) \geq T \\ 0 & , \quad \Delta_n(u,v) < T \end{cases} \quad (2)$$

Temporal differencing is a very simple technique to implement and does not requires large computational resources. This technique can be used in systems with high temporal requirements, but not systems that present requirements for the detection of objects with extremely precision, becomes impossible the utilization of a camera motion [13]. Since, if the object stops moving, or the light in the environment changes this method will bel unsuccessful and requiring the human intervention to adjust the thresholder value. Besides that, the adding of undesirable background regions on the periphery of the target, where the object has "just been" [12]. Although this problem could be fixed using the knowledge of the target's motion this will decrease its efficiency, not being suitable for real video surveillance situation.

2.2.2 Optical flow

Instead of first finding the pose of the human in some discrete frames and then using this information to calculate the motion, one may measure the human motion between images and set up an inverse kinematic framework which makes it possible to calculate the corresponding motion in the 3D model, called the model, therefore the poses are updated based on the motion in the images. This was first done

by Yamamoto and Koshikawa [6], in 1991. They measured optical flow within various body parts and used that through a Jacobian matrix to update the model.

Ju *et al.* [14] used two planar patches to model a leg, creating a cardboard model. The motion of each patch is defined by eight parameters. For each frame the eight parameters are estimated by applying the optical flow constraint on all pixels in the predicted patches. The distance between the corners of the predicted patches are constrained to reduce the complexity of the estimation. Note that this concept will be further discussed below.

Bregler and Malik [15] extended the concept of Ju *et. al.*[14] by introducing a twist motion model and exponential maps which simplify the relation between image motion and model motion, having the advantage of being open to both single and multiple views.

Optical flow or optic flow has been projected as a preprocessing stage for a several high level vision algorithms [16].In a visual scene, this method uses the relative motion between an observer, in this case a camera, and the scene. In other words, it analyses images in a sequence to check, for each point, the detected motion. With the utilization of optical flow, the movement of all points of the image is represented by a field of vectors that indicates the velocities or discrete of the image displacements of a certain point. Optical flow has as main advantage the possibility of segmentation of objects in motion, such as in autonomous navigation. However, in this case it becomes necessary to eliminate the noise from the optical flow induced by the camera motion [17].

Although this technique has developed huge progresses, the calculation of the optical flow is a task that requires enormous computation resources as well as a special image processor for this purpose. This is because of the need to process images separated by short periods of time, usually forty milliseconds. In addition, it assumes that all the parts constituents of the human body moved together in the same direction [9], making this solution not suitable for the detection and tracking of humans.

Another problem by using optical flow is the need of a large number of filters, becoming very expensive and very sensitive to temporal aliasing in the image sequences [18].

In an outside environment, a system for tracking human motion must be prepared for variations in terms of lightning, such as clouds attenuating the sunlight on objects in the scene. However, this solution assumes a constant illumination over time, rendering it unfit for an outside environment application.

2.2.3 Background subtraction

Background subtraction has been commonly used for detecting moving targets, in video, from static cameras [19]. In this approach, there are two types of frames, the current frame and the reference frame

or “background image”, Figure 2. The main concept of this technique is to subtract the image from a reference image that models the background scene. It can be explained into three steps. The first step, is called Background modeling, constructs a reference image representing the background.

The reference frame is taken when there is no activity in the scene, to make possible that this method is efficient, this image should be updated frequently to adapt to changes of lightning, weather and objects being introduced or removed from the scene. Overtime time, if there was no re-initialization, the errors would accumulate, turning this technique unappropriated for tracking a scene that has significant variations [20].

The second step determinates an appropriate thresholder value to subtract needed to achieve a desired detection rate, Threshold selection. Finally, the third step involves pixel classification to categorize the typo of a given pixel, for example, the pixel that is part of the background or the moving object.

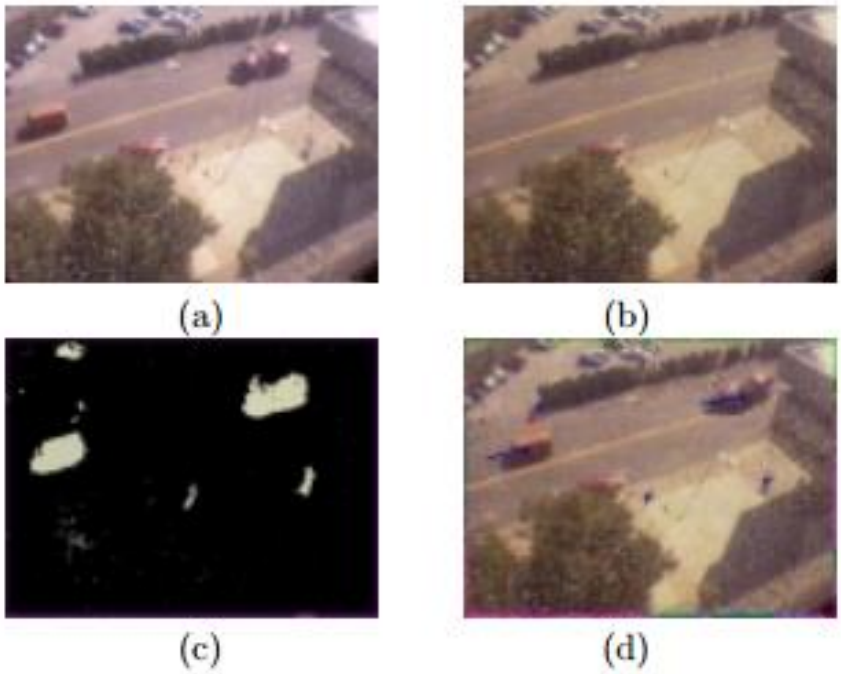


Figure 2 - Background subtraction[20]

Like the optical flow technique, the background technique has several approaches, such as, Gaussian distribution, mixture of gaussians (MOG), multibackground. These approaches differ in the type of the representation model of the background model, how the motion identification is processed at each point and the strategy used to update the background model.

The background model using Gaussian distribution involves the calculation of two masks for define an average value and respective and its variation. Also, requires, when the system is initialized, a learning reference image. This technique uses a multiple of the standard deviation to determinate the threshold value. In this way, it becomes more sensitivity of the detection adjusted to the variation of the brightness

and color, at each point. However, this technique is not efficiency, because it cannot detect all variations that occur on the background. Such as, the existence noise and certain movements of objects belonging to the background. For that reason, it was proposed another model. This model was called Gaussian mixture model. Although, significant progress has been made since the paper of Satuffer et. al.[20], regarding the problem of high computational time, there is not completed improved to deal with the constraints of real-time application [21].

Another approach, consists of the use of multibackground develop by Boult et. al. [17], in 1998. In this approach the backgrounds are defined by a model of conditional increment, not calculating the point variance. Instead, a gray point measurement of the point value to determine which of the backgrounds is closer to the present value. When updating the multiple reference images, the threshold of each point is also updated. Thus, for each point, this value is incremented or decremented depending on the region that is detected. However, this dynamic threshold update procedure can cause stability in the segmentation process.

In a background subtraction approaches the major disadvantage is inefficient in cases that the background does not have a significant visible portion. In scenes with lots of moving objects, especially if they are moving very slow, the algorithm works improperly [20].

2.3 Image features models

This section discusses the different kind of image features that can be extracted from a video sequence to represent the actions spatial structure.

2.3.1 Body models

In action recognition, there are a lot of model types, for example, body models, image models and local statistics models that detect human body motion. Although, all of them are distinguished by the amount of high level information processing abilities versus their efficiency.

Body models consists of observing in all video frames the pose of a human body and recover a variety of available image features, performing action recognition based on such pose estimates [2].

In 1978, Marr and Nishihara proposed a representation of a three dimensional body model consisting of a hierarchy of 3D cylindrical primitives [22]. Such model was later adopted by several studied approaches. Later more flexible body models based on super-quadrics have been used, as well as models based on a textured spline.

The approaches start from tracked patches in 2D and then lifts the 2D configurations into 3D, Figure 3, however, 3D pose reconstruction from a single viewpoint is a challenging problem, because of the large number of parameters and the ambiguity caused by perspective projection [23].

Motion capture techniques requiring special markers attached to the human have also been used for action recognition. Other approaches directly work on the trajectories of 3D anatomical landmarks, for example, head and hand trajectories [2].

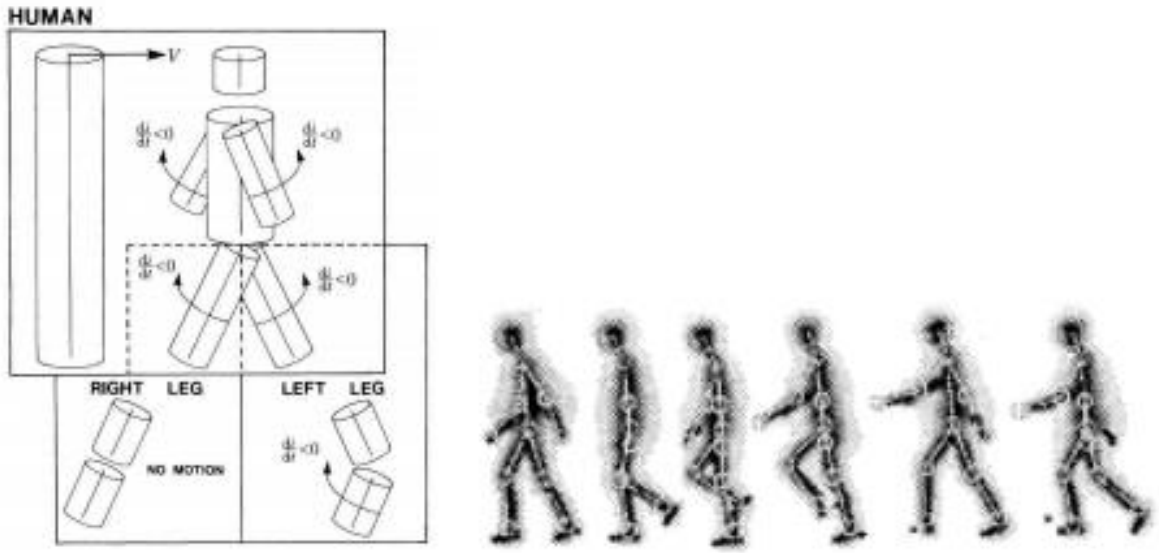


Figure 3 - Differences between 3D cylindrical primitives and 2D tracking techniques [2].

Direct recognition approaches work from 2D models of the human body, without lifting these into 3D. Other direct recognition approaches use coarse 2D body representations based on tracked blobs and patches, e.g. hand and head trajectories.

However, it is important to note that finding body parts and estimating parametric body models from images remains an unresolved problem, independent of the model used (2D or 3D) [2].

2.3.2 Image models

Image models use a region of interest (ROI) instead of detecting some body parts, like in body models. Usually this kind of approach detects silhouettes and contours of the human body that is performing the action. In this case, this models can be more efficient in some situations, however, if the background is not a static black background it needs to have a pre-processing stage, background segmentation. Although, in the last few years, there is an increasing use of image models that uses dense optical flow extracted from consecutive images [2].

The advantages of image models is that doesn't require background models, so there is no need of using background subtraction, reducing the overall processing time. However, problems with the lightning and the changes in the material properties may affect the result of this strategy implementation.
Another imperfect class of image models is the class of image features based on gradients. Its greatest disadvantage subsists in that static nonmoving body parts can be easily confused with static objects in the background with strong gradients.
Overall, this type of image models class may be simpler comparing to the body models, although, they are very sensitive to variations of light, gradients and the material properties. To attempt to resolve this problem, there were some combinations of the two different classes, such as gradients and flows, or silhouettes and flow, resulting in an increasing of the successful result rates.

2.3.3 Spatial statistics

Spatial statistics uses small regions of an image or video to analyze, therefore, there is no need to individually identify a body part.
Space time interest points were used to generalize interest points and local descriptors. Such approaches are typically based on bottom up strategies, which first detect interest points in the image, mostly at corner or blob like structures, and then assign each region to a set of preselected vocabulary-features. Image classification reduces then to computations on so called bag of features, BOF, for example, histograms that count the occurrence of the vocabulary-features within an image.
In summary, statistical methods based on local features have recently drawn a lot of attention in the action recognition community because they promise the same advantages as in static object recognition, and because they can easily be applied to difficult scenes, e.g. movies or video clips from the internet, that evidently will be very difficult to model with full-fledged image or body models. However, the very nature of complex human actions will probably make it necessary to combine those methods with stronger spatial and temporal models, e.g. computing spatial statistics over dense (rather than sparse) image grids, and relying on human detection for scenes containing multiple persons.

2.4 Machine Learning

Arthur Samuel affirms that machine learning "gives computers the ability to learn without being explicitly programmed" [24]. In other words, machine learning explores the study and construction of algorithms that can learn from and make predictions on data, such algorithms overcome following strictly static

program instructions by making data-driven predictions or decisions, through building a model from sample inputs [25]. In action recognition systems, the machine learning algorithms are used for classify the actions of the human in a video surveillance, as seen in Figure 4. In other words, the main objective of the machine learning algorithms is to find the most probable action according to the parameters. So, it has to estimate which posture the current image stands for, then recognize which action the posture sequence means.

Figure 4 – Examples of different types of human actions classification [26].

The most common question when data is already determinate is what type of analysis it is need. So, the answer can be based on four solutions:

- Regression:
    - Predict new values based on the past, inference;
    - Compute the new values for a dependent variable based on the values of one or more measured attributes;
- Classification:
    - Divide samples in classes;
    - Use a trained set of previously labeled data;
- Clustering:
    - Partitioning of a data set into subsets, clusters, so that data in each subset ideally share some common characteristics;

Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is feasible; example applications include spam filtering, optical character recognition, OCR[1], search engines and computer vision.

[1] Optical character recognition, OCR is a technology for recognizing characters from an image or bitmap file whether they are scanned, handwritten, typed or printed. In this way, through OCR it is possible to obtain a text file editable by a computer.

There are many types of algorithms, such as Bayesian Network (BN), Hidden Markov Model, Naïve Bayes, Bayesian Belief Network (BBN), among others, however in this section, it will be discussed some machine learning algorithms to better understand some human action recognition systems, that it will be studied in section 2.5.

2.4.1 Support Vector Machine

In machine learning, there is a task of inferring a function from labeled training data, called predictive or supervised learning. In supervised learning, the main objective is to learn a mapping from inputs, x, to outputs, y, given a labeled set of input-output pairs $D = \{(x_i, y_i)\}_{i=1}^{N}$., where D is the training set, and N is the number of training examples [27]. In each example is a pair consisting of an input object, normally a vector of numbers, like for example to represent the height and weight of a person, called features, attributes or covariates. The output or response variable is usually a categorical or nominal variable from some finite set, $y_i \in \{1, \dots, C\}$, for example as male or female, or a real-valued scalar [27].

These methods are usually fast, accurate and most widely used in practice. One example of supervised learning model is support vector machine or it can be also called support vector networks. This model is mostly used for analyzing data for classification and regression analysis [28].

A support vector machine model, SVM, is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible, as shown in Figure 5. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [28].

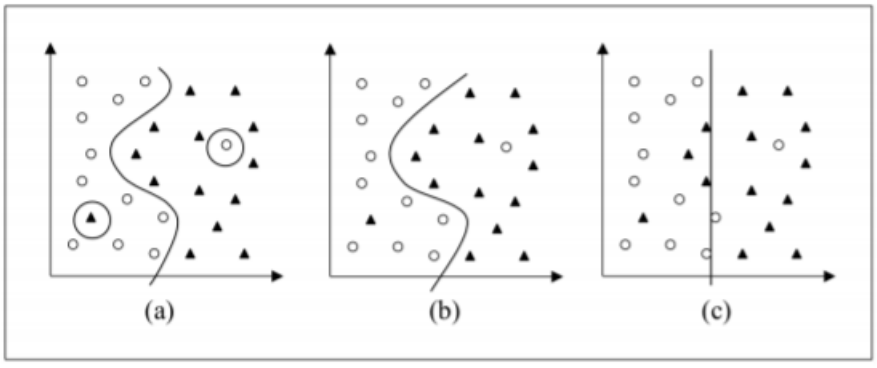


Figure 5 – Example of the Support Vector Machine model by using three hypotheses [28].

2.4.2 Hidden Markov model

A Hidden Markov model, HHM, was first used by Yamato *et al.* [29] in action recognition. In their work, they used Hidden Markov based methodology for learning actions, such as, to recognize six tennis strokes among three players, Figure 6. The name of this model comes from two properties:

- Assuming that the observation, in a certain time, was generated by a process whose state is hidden from the observer;
- Assuming that the state of this hidden process satisfies the Markov property;

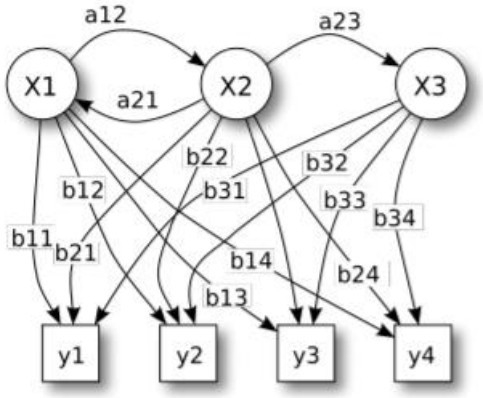


Figure 6 – Representation of HMM using three states [30].

Markov property affirms that given the value of S_{t-1}, the current state $S_t$is independent of all the states prior to $t-1$. So, the state at some time encapsulates all the information about the history of the process, in order to predict the future of the process.

Once a system can be described as a HMM, three problems can be solved: evaluation, decoding and learning. The first two are pattern recognition problems: finding the probability of an observed sequence given a HMM (evaluation); finding the sequence of hidden states that most probably generated an observed sequence (decoding); the third problem evolves a generating a HMM given a sequence of observations (learning).

2.4.3 Maximum a Posteriori

Maximum a Posteriori, MAP, estimate a priori knowledge of an unknown quantity that equals the mode of the posterior distribution. Also, MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data.

It derives from the Maximum Likelihood Estimation, MLE. These two approaches are basic principles for learning parametric distributions. However, MLE has limitations in estimate a robust model parameters

when there are only a few training data available. Therefore, the difference between MLE and MAP training is in the definition of the prior distribution for the model parameters to be estimated. So, in order to obtain a more precise model it has to be used the MAP.

2.5 Algorithms for analyzing human motion

In section 2.3 was presented the different kind of image features that can be extracted from a video sequence to represent the spatial structure of actions. Therefore, this section discusses some techniques that try to extract some particularities, for later to be used on analyzing their actions over time. For example, in Figure 7, after the moving target is already detected and some anomalies are eliminated, by the pre-processing stage, it is possible to proceed to human features extraction.

As show in Table 1, there is a various range of ways to compare image data, from edges, silhouettes, blobs, depth and many other ways to try to extract information about a human action.

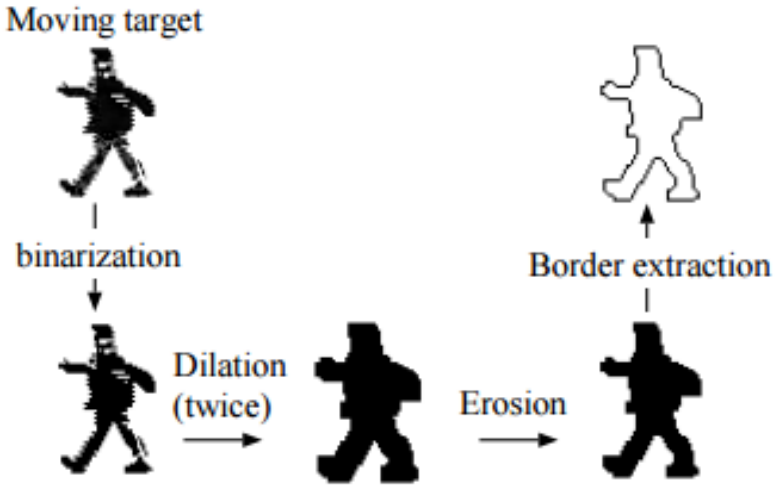


Figure 7 - Real-time target extraction then clean the errors so finally it is possible to proceed to the extraction.

Table 1- Papers where a human model is used directly [6]

First author	**Model type**	**Parts**	**Object**	**Abstraction level**	**Dimension**[2]
Hogg [31]	Cylinders	14	Body	Edges	$2\frac{1}{2}$
Lerasle et. al.[32]	CAD Model	2	Leg	Texture	3
Meyer [33]	Boxes	6	Body	Contours	$2\frac{1}{2}$
Munkelt [34]	Modified Cylinders	10	Body	Joints	3
Yaniz [35]	Stick-figure	16	Body	Sticks	3
Hu [36]	Rectangles	10	Body	Silhouette	2
Plänkers [37]	Ellipsoids	6	Arm	Depth	3
Moeslund [6]	Cylinders	2	Arm	Silhouette	3
Rosales [2]	Cylinders	10	Body	Silhouette	2
Wren [38]	Stick-Figure	5	Upper body	Blobs	3
Yamamoto [6]	CAD Model	11	Body	Motion	3

In 1983, Hogg wrote the first analysis-by-synthesis publications, a model for detecting a human motion by using image subtraction to obtain a boundary box of a human, and then he compared the edges of the box with edges expected from a human [31].

Unlike edges models, silhouette models are a region-based data type, so it presents the advantage of being less sensitive to noise. On the other hand, some details may be lost in the extraction of the silhouette [6], this topic will be discussed in more detail below.

In 1997 a contour model was introduced by *Meyer et al.* [33] detecting the parts of the human from optical flow and represented by their contours. Then, the contour will be compared to some predicted model contours. The problem of using contour models is rise of computation time.

[2] Dimension can be: 3D, 2D and $2\frac{1}{2}$D which refers to a 3D pose data based on 2D processing or testing a 3D pose estimating framework on pseudo-3D data.

Another representation of the human body, it can be achieved by its joints or it stick-figure, since it reflects the physical structure of the human body, as shown in Figure 8. These two representations are closely related, however the stick-figure is compared to an image skeleton found from the silhouette.

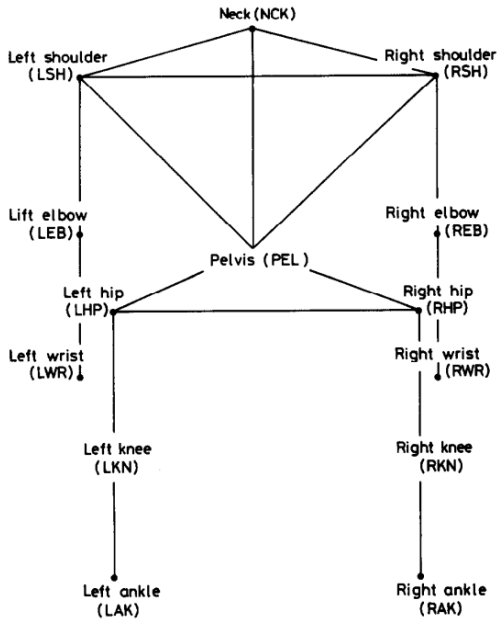


Figure 8 – The stick-figure of a human body: 14 joints and 17 segments [39]

Lee and Chen [39], proposed an interpretation of a specific 3D body structures of a human, by having a set of 2D projected coordinates of the joints as an input data . In this case, they used the six features points on the head to determine the camera position: neck, nose, right eye, left eye, right ear, and left ear. Then, it is estimated the stick figure model by 14 joints and 17 segments. They also assume that the head points and the 14 joints are already identified on the image plane beforehand [39], making this part an issue for a surveillance video in an outside environment.

Another abstraction level with issues in surveillance video, in an outside environment, is called depth data, first introduced by Plänkers, in 1999 [37]. The most renowned example is the data depth of a Kinect sensor, using 25 joints of information which is popular in motion sensing games.

An alternative to the representation of human action is texture representation. This technique was presented by Lerasle *et al.* [6], in their work they generate a 3D model of a human's leg, and during the process it was compared the texture of the image with the texture of the model using correlation [6].

Another abstraction level was used by Wren and Pentland [2], it was called blobs detection. In computer vision, blob detection methods are aimed at detecting regions in a digital image that differ in properties, such as brightness or color, compared to surrounding regions. Informally, a blob is a region of an image in which some properties are approximately constant, and all the points in a blob can be similar to each other.

A complementary information about certain regions could be easily obtained by the use of blob detection, instead of edge detectors.

In the work of Wren and Pentland, called *Pfinder* [3], blob detection was used to obtain regions of interest for further processing. However, their work will be further discussed in below.

2.5.1 Person Finder

Pfinder, or "Person Finder", is a real-time system that detects features in video images in to recognize human figures and their movements and gestures. It uses a multi-class statistical model of color and shape to obtain a 2D representation of the human features [3], by identifying the boundaries of a person in the image, it also analyses the regions inside boundaries and relates them to the known structure of the human body. As mentioned above, this system uses a 2D blobs, Figure 9, for the head, the hands and feet, as the abstraction level. When a hand or foot is occluded, for example, it is deleted from the human model. However, when it reappears, a new blob is created.[2].

Figure 9 – (Left) it is the video input, (right) a blob representation in 2D.

This algorithm has been used in various applications, such as, wireless interfaces, video databases, video games [40], and gesture recognition systems (American Sign Language) [41].

To detect movement this algorithm acquires a sequence of video frames trying to learn the scene, getting information on the color covariance associated with each image pixel of the scene. After the background is complete detected, if *Pfinder* notices variations, it will form the 2D human model, like it shows in Figure 10.

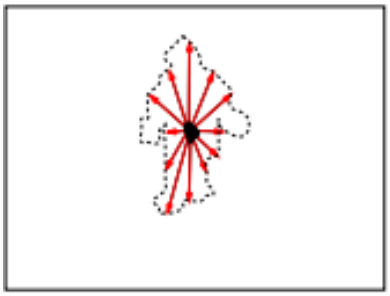


Figure 10- Building 2D person model [3]

In terms of limitations this model only recognizes one human in each scene, so if there is a crowd in the image, the system will attempt to analyze them as one single human figure. Besides that, if there is an increase of significant changes in the scene, like for example in the light, the *Pfinder* will not be capable of working properly, because it will try to incorporate the changes in the scene into the human figure model. Furthermore, this algorithm does not cope with multi backgrounds, because the histogram of the pixel intensity will contain more than one distinct peak[3].

Although, there are several issues that still need to be addressed, this algorithm presents an enough stable solution to support real time applications as well as higher-order vision techniques for an unchanging background, however, it should be improved if the purpose of this algorithm is to be used in an outside environment.

2.5.2 W^4

W^4 (Who? When? Where? What?) is a real time surveillance system for detection of human actions in monochromatic video sources [4]. This system presents an advantage of detecting multiple humans in the same image

This system works using a combination of shape analysis and tracking to detect the human features (head, hands, feet and torso), without the use of color cues, and detects at the same time interactions between human and objects (i.e. people exchanging objects, leaving objects in the scene, taking objects from the scene, among others).

Its applications are more indicated for an outdoor surveillance[3], with a stationary camera, in particularly for night-time, because it supports a low light condition environment, as it shows in the Figure 11.

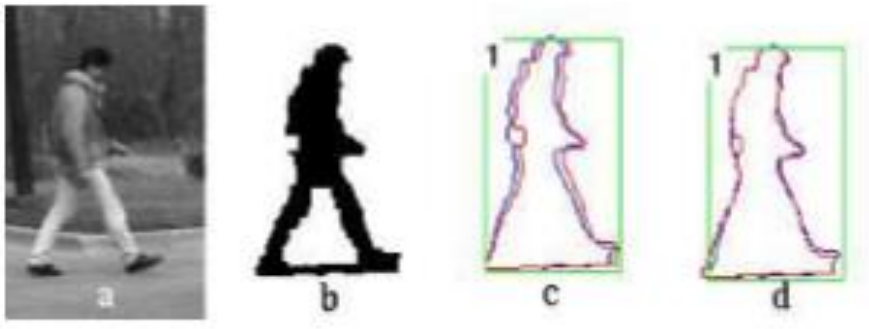


Figure 11 - Detection of movement a: Input image; b: background deleted; c:silhouette edges based on difference in median; d: final alignment after silhouette correlation [4]

[3] W^4 system has also been used in an indoor environment, but in this case it uses *Kalman* filters and kinematic constraints [38].

Identically to *Pfinder*, this system uses background subtraction. In other words, this system classifies each pixel into three categories: minimum intensity values (M), maximum intensity values (N) and the maximum intensity difference between frames (D) [4]. So, using the background model, it is possible to classify pixels in a background or foreground pixel. Therefore, in an image *I*, the pixel *x* is a foreground pixel if [4]:

$$|M(x) - I(x)| > D(x) \quad \text{or} \quad |N(x) - I(x)| > D(x) \quad (3)$$

Besides that, W^4 uses thresholding on some operations to decrease the level of noise, possibly caused by illumination changes. After that, some features of the foreground object are characterized for a label, such as, the median, the centroid and the bounding box, as show in Figure 12 a). At the end, this system uses an existing model, *Cardboard People Model,* to track and recognize the body parts, Figure 12 b).

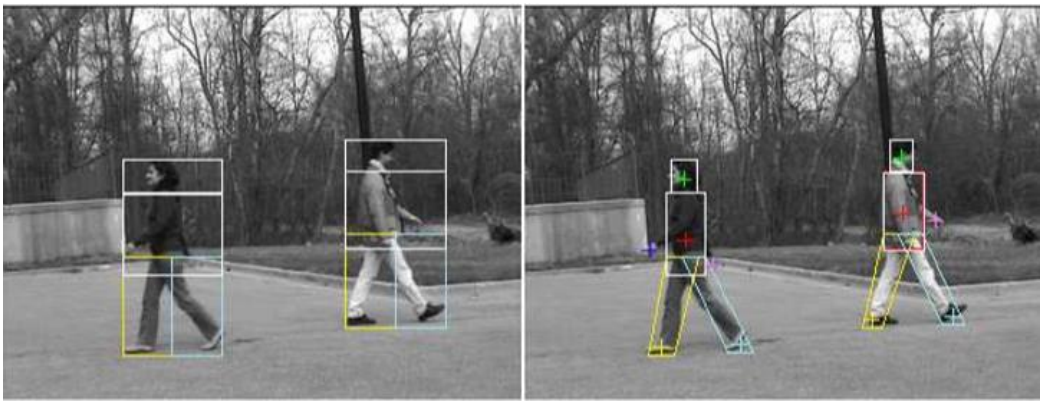


Figure 12 - Detection of foreground regions using Bounding boxes (a) Utilization of cardboard model (b) [4]

As with all systems presented there are some points that need to be improved. In case of W^4, it can only be used for upright people and not for other types of human poses. Besides, this system does not detect shadows in the images. At first sight, this could not be interpreted as a big problem, however, in a real situation, where the shadows are a huge factor for a bad result in the process of segmentation this limitation can decrease the system efficiency. For example, if there is a group of people in a scene, the intersection of them could not be noticed by the system, because they will have a similar hue in a grey scale, becoming impossible to identify interactions between humans [4].

As mentioned above, W^4 uses the *Cardboard People Model* to represent some of the human body parts[14]. Here the body parts are viewed similarly with planar regions that have connected patches as the articulations of the body.

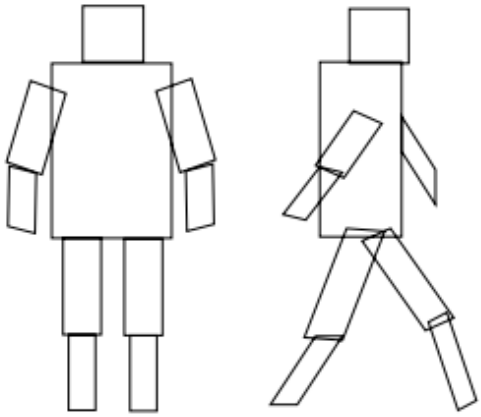


Figure 13- The Cardboard Person Model [14].

In this case, to build a chain structure, each patch is linked to just one preceding patch and to a following patch. As we can see in the Figure 13, there are four corners for each patch, and two of this corners will be defined as the articulating points. Therefore, each patch shares a mutual edge with another, making the person body structure [14]. Although, this model presents a possible method for tracking human motion, it has some issues. For example, it doesn't work properly in a long distance motion, because the regions for tracking are very small. The "motion" of the clothes can also interfere when the model is trying to track all the patches and this model.

2.5.3 Microsoft Kinect

The Kinect is a motion sensor developed by Microsoft, introduced in November 2010, Kinect 1.0 (Figure 14). Initially, it was created as an input device for the Xbox 360 gaming console and currently optimized for their operating system, Windows.

Figure 14 – Microsoft Kinect 1.0 [42]

The data acquisition process is called RGB-D and can be understood as the uptake of an image with color, measuring its depth being held, with light techniques structure.

Its main function is the production of three-dimensional data in all types of light conditions. However, it also executes human activity analysis and hand gesture analysis. Therefore, a rough skeleton of a person can be easily obtained, by using a Microsoft Kinect sensor, as shown in Figure 15.

Figure 15 – Example of an approach for detection of a human skeleton[4] by using the Microsoft Kinect sensor [43].

This has resulted in renewed interest towards increased research on skeletal features for human motion representation. Several new datasets have provided researchers with the opportunity to design novel representations and algorithms and test them on a much larger number of sequences. Recently the focus has shifted towards modeling the motion of individual joints or combinations of joints that discriminate between actions.

Ofli et al. [44] proposed the Sequence of Most Informative Joints, SMIJ, representation, a new and highly interpretive feature for human motion representation for skeletal data based on joint angle time series. Wang et al. [45] proposed a feature mining approach for computing discriminative action using a recursively defined temporal pyramid of joint configurations. However, nowadays the Kinect sensor is, in general, an RGB camera, a depth sensor, a set of microphones and an accelerometer, as can be observed in Figure 16, considering an inexpensive system.

[4] In this approach, the skeleton model is controlled by 32 degrees of freedom grouped in 9 joints., without including any type of detail about the hands, the head and the feet [43].

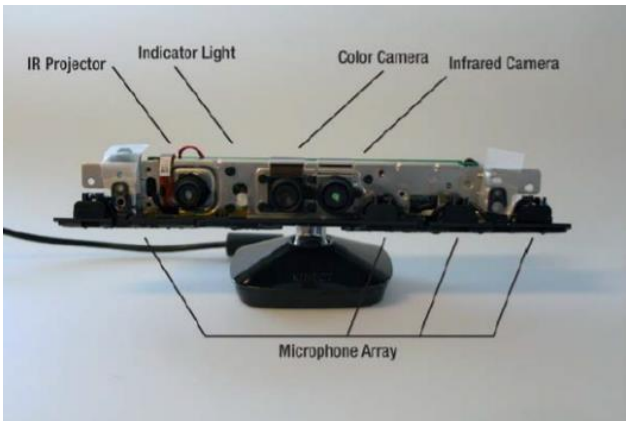


Figure 16 – Kinect for Windows components [46].

In Figure 16, it is possible to see the specifications of the Kinect 1.0. This camera supports a resolution of 640 x 480 pixels 30fps, up to 1280 x 960 pixels at 12fps. The camera used for depth measurement is infrared and supports a maximum resolution of 640 x 480 pixels at 30fps [46].

The depth sensor contains a combined infrared light projector with an infrared camera, which is a Complementary Metal-Oxide -Semiconductor, CMOS, monochrome able to obtain a three-dimensional Model. The light emitted by the infrared projector is passed one diffraction grating, which can cause the light emitted from becoming small dots that after are captured by the camera. Therefore, the depth is measured by the distance between the sensor and human [46].

The Kinect used four microphones with echo cancellation and noise suppression, to achieve a capturing sound that involves the whole environment and obtains a direction of the sound source. There is also a 3-axis accelerometer configured for a 2G variation, where G is the acceleration due to gravity, that allows to determine at each instant the Kinect sensor orientation [46].

Despite the use of Microsoft Kinect sensor has make a great incentive for the development of many projects in many areas, like for example, in robotics, computational science, electronic engineering and medicine, Kinect also features a field of view in pyramid form that incorporates some limitations. Allowing the recognition of objects or users more precisely between the 40cm and 4m, with a viewing angle of 57 degrees horizontally, 43 vertically, as show in Figure 17 [46]. This distance limitations makes a huge disadvantage of the Microsoft Kinect sensor for a human recognition in a real outside environment.

Table 2 - Kinect 1.0 specifications [47].

Energy consumption	2.25W
Distance of use	0.5m to 4.5m
Framerate	Approximately 30 Hz
Angular Field-of-View	57° Horizontal
	43° Vertical
Sensor	RGB
	Depth
	2 * Microphone
Nominal spatial range	VGA (640x480): 30fps
Nominal spatial Resolution (at 2m distance)	3 mm
Nominal depth resolution (at 2m distance)	1 cm
Device connection type	USB
Operation environment	Inside
Dimensions	24.9 x 6.6 x 6.7 cm
Price	147.09 (Amazon)

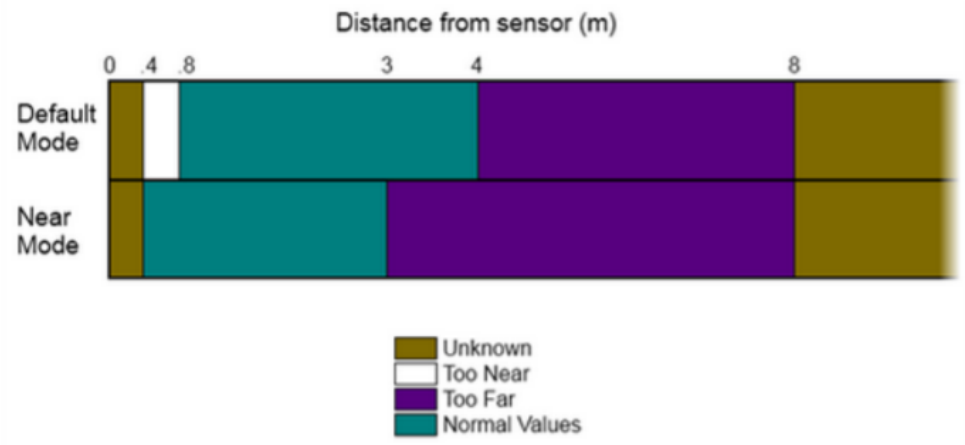


Figure 17 – Field of the Microsoft Kinect sensor [48].

2.5.4 Star skeleton

Normally, virtual human bodies are structured as articulated bodies defined by a skeleton. A skeleton is a connected set of segments, corresponding to limbs and joints. A joint is the intersection of two segments, which means it is a skeleton point where the limb linked to that point may move.

Star skeleton is a technique to track mainly human targets, using their skeleton to detect and create a human contour, by analyzing their broad internal motion features, Figure 18. Others techniques, such as W^4, tried to use whole human contour, however, their strategy proved to be less efficient to implement than this technique since each border point is very similar to its neighbor points [49].

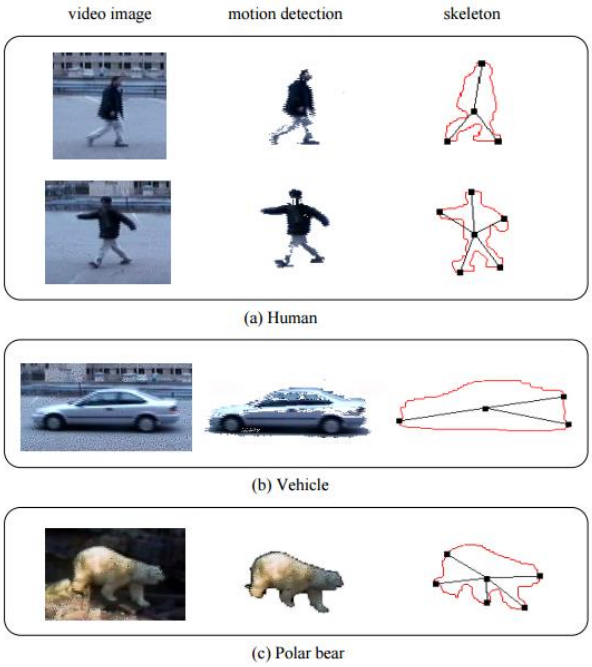


Figure 18 - Several types of moving targets and theirs related skeleton[5]

This approach, can possibly be applied to other objects, such as animals and vehicles, as we can see in (eq. 3) and (eq. 4). So to detect the star skeleton of the object it is necessary to determinate the centroid of the object first, using [49]:

$$X = \frac{1}{N_a} \Sigma_{i=1}^{N_a} X_i \qquad (3)$$

$$Y = \frac{1}{N_a} \Sigma_{i=1}^{N_a} Y_i \qquad (4)$$

In these two equations, (X, Y) is the average boundary pixel position, N_a is the number of boundary pixels, (X_i, Y_i) represents the pixel on the boundary of the target.

To calculate the distance between the centroid (X, Y) to each border point (X_i, Y_i), it should use:

$$D = \sqrt{(X_i - X)^2 + (Y_i - Y)^2} \quad (5)$$

Besides that, for reducing the noise, it must calculate D_i (dimensional discrete function of D), to find the zero-crossing of the difference function local maxima.

Star skeleton presents some advantages compared with others techniques. This kind of real-world implementations have to be computationally inexpensive and be applicable to real environments in which targets are small and data shows a lot of noise [5].

Star skeleton is a fast skeletonization technique achieved by connecting centroid of the target object to contour extremes. Therefore, the utilization of just several pixels to detect human action motion, such as, body inclination, walking and running, turns this technique more efficient. Despite all this advantages, there is still a lot to develop, like for example in terms of complexity of the human's motions and then apply it multiple objects.

2.6 Conclusion

In this section, it will present the conclusions as well as a brief summary of the human action recognition systems studied in section 2.5, as shown in Table 3.

Table 3 – Characteristics of the systems.

System	**Human features extractions**	**Segmentation model**	**Classification model**
Pfinder [3]	Blob (2D representation)	Background subtraction [15]	Maximum a Posteriori Probability (MAP) [3]
W^4 [4]	Bounding box (Detection of the head, hands, feet and torso)	Background subtraction [4]	The Cardboard model [14]
Microsoft Kinect [50]	Joint locations (Skeleton representation)	(Depth Sensor [5]) [47]	Hidden Markov Model[6] [51]

[5] Microsoft Kinect does not use an object segmentation to track human movements, but instead uses the depth sensor, as shown in section 2.5.3.

[6] There are several studies that use another classify model, but in this system, it was used the Hidden Markov Model.

Star Skeleton [49]	Border extraction	Background subtraction [52]	Support Vector Machine [53]

The video surveillance aims to prevent certain situations that can for example degenerate into crimes, by detecting suspicious activity, causing alarm or similar actions. In this case, systems capable of targeting, identification and tracking human motion became indispensable to reduce the probability of these situations to happen. However, action recognition remains a challenging problem, and so far, only the tip of the iceberg has been revealed. Most of the existing solutions only work for simple or synthetic scenes which have been created for testing action recognition algorithms.

The best solutions still require large amounts of supervised training data, which is expensive to obtain. Thereby rendering it very difficult to compare and grade the different systems, based on various test data and assumptions. Still, there are three concepts which may be considered the main performance parameters in any motion capture system: robustness, accuracy, and speed [2].

Surveillance systems are aiming at very robust performance since they will often be working continuously, autonomously for a long period and in uncontrolled locations. So, it should not be sensitive to any kind of changes, such as, lighting, weather, the number of people in the scene, clothes, etc.

The next concept, accuracy, refers to the similarly between the captured motion and the actual motion. The most of the times, this concept is directly proportional to the size of the human in the scene.

The processing speed of the system is usually calculated by the time of each frame is processed before a new frame is recorded, being divided into real-time and offline processing.

3. SYSTEM IMPLEMENTATION

In this dissertation, it is proposed a new algorithm for human action recognition. Therefore, in this chapter, it will be discussed the algorithm and the advantages and disadvantages comparing with others algorithms that were discussed in the second chapter.

All experiments were conducted using a common Asus laptop with 2.59 GHz Intel Core i7 processor.

3.1 Human action recognition system

Human action recognition aims to recognize different actions from a series of observations on the agents' actions and the environmental conditions. This research field has captured the attention of several computer science communities due to its strength in providing personalized support for many different applications and its connection to many different fields of study such as medicine, human computer interaction and security.

A human action recognition system contains several steps, such as, object segmentation, posture extraction and machine learning model. In this dissertation, the object segmentation consists of the background subtraction, the pre-processing stage and the blob detection, as shown in Figure 19.

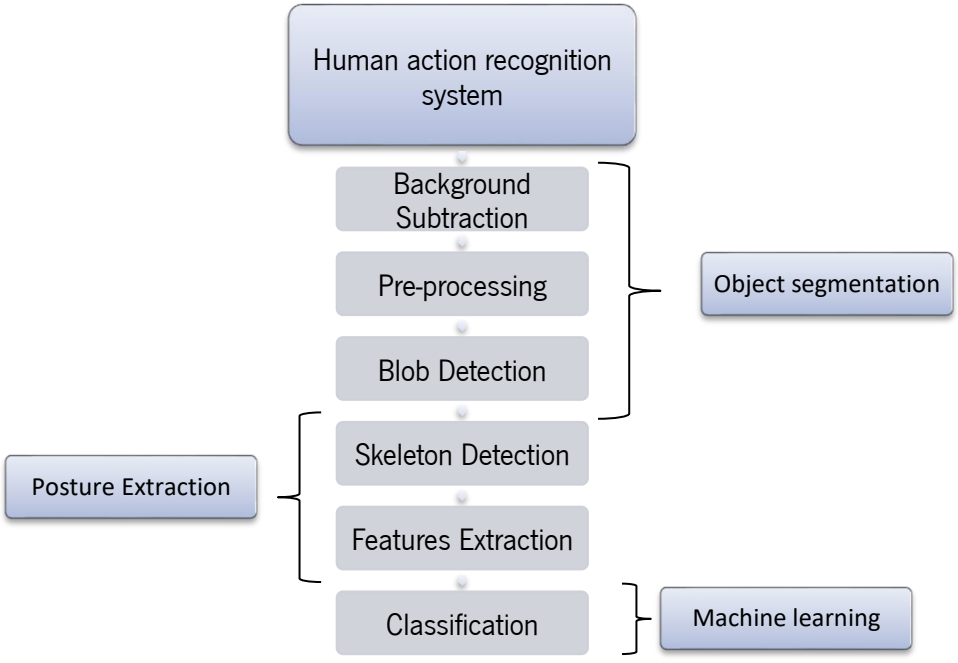


Figure 19 – Human action recognition system.

These three stages are inserted in the object segmentation algorithm. The next step, posture extraction, involves a skeleton detection and the physical features extraction. Finally, there is the machine learning model that corresponds to classification of the human action.

These steps had to be considered, in order to develop this work. The main objective of this dissertation is to test the feasibility of using detection of human features in a skeleton silhouette for detecting human actions in a video surveillance, as show in Figure 20, in order to contribute for the development of a human action recognition system that could be more effective than the systems presented in section 2.5.

Figure 20 – Video surveillance in an outside environment, motion detection.

3.2 Software

For the development of this dissertation the Matlab® software[7] (version 2016a) was selected and it was built around the MATLAB scripting language. A software with multi-paradigm numerical computation environment and fourth generation programing language, developed by MathWorks, in 1984 [54].

MATLAB® was first used by researchers and practitioners in control engineering, but quickly spread to many others domains [1]. Nowadays, it is also used in education, the teaching of linear algebra, numerical analysis, and is popular amongst scientists involved in image processing, because this software has a very large and growing database of built-in algorithms for image processing and computer vision applications [54]. Besides, this software can read in a wide variety of both common and domain-specific image formats. The Matlab Desktop environment allows to work interactively with the data, helping to keep track of files and variables, and simplifies common programing/debugging tasks [54].

Another advantage for using MATLAB® is the Computer Vision System Toolbox™ that comes with this software [55]. Computer Vision System Toolbox™ provides algorithms, functions, and apps for designing and simulating computer vision and video processing systems, such as, feature detection, extraction, and matching; object detection and tracking; motion estimation; and video processing. Furthermore, for 3-D

[7] http://www.mathworks.com/products/matlab

computer vision, the system toolbox supports camera calibration, stereo vision, 3-D reconstruction, and 3-D point cloud processing [55].

3.3 Implementation

Detection of moving objects and motion-based tracking are important components of many computer vision applications, such as activity recognition, traffic monitoring, and automotive safety. The problem of motion based object tracking can be divided into two steps:

1. Detecting moving objects in each image in the video surveillance;
2. Associating the detections corresponding to the same object over time.

Digital video consists of the sequence of images, frames, in a fixed and sufficiently short times, in order to create in the observers the perception of uninterrupted movement, as shown in Figure 21 [17].

Figure 21 – Sequence of images from a video [17].

This work used a video of a real-word dataset created in University of Minho, during a total of 7 days, in which presents an indoor and an outdoor environment. This dataset was captured by two cameras, HIK Vision and IR Network, as shown in Figure 22

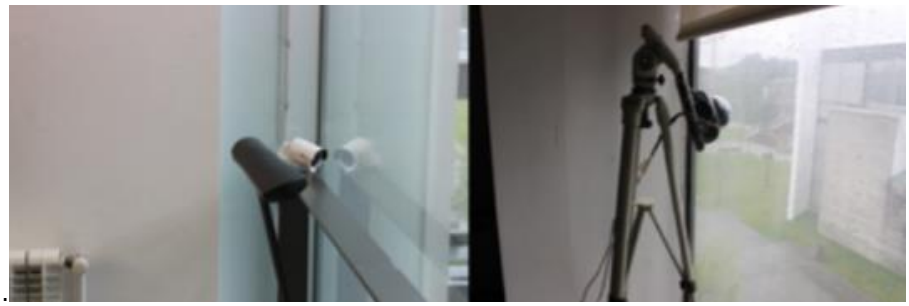


Figure 22 – HIK Vision and IR Network cameras [30].

The dataset presents some different conditions, especially in the outdoor environment, because of the weather, mostly, rainy and windy days, making it difficult for the detection of movement. This problem comes from the use of some objects, like for example, umbrellas, heavy coats, among others. As the detection was not controlled, being real-life moments in a university environment, Figure 23, it becomes possible to test the efficiency of the respective algorithm.

Figure 23 – Example of images collected [30].

3.3.1 Background subtraction

Object segmentation is the detection of moving objects in sequences of video images [7]. By the analyzing of the results from the object segmentation that makes it feasible to monitor human activity through the people in a scene. In other words, with the segmentation of moving objects, it is expected to define the certain regions of points of the image to describe the shape and color components of all objects not belonging to the background of the scene, as shown in Figure 24. The ideal algorithm for segmentation should be sufficiently robust against environmental variations, but at the same time sensitive enough to detect the moving objects of interest.

After studying the different types of object segmentation in section 2.2, it was concluded that, the best model is the background subtraction approach. It is the technique that best adapts to the application in question, in particularly due to its performance. Therefore, for this work it was used an algorithm that was previously implemented by Duarte Duque (2008) [17]. This algorithm is capable to be applied in real-life situations, for example, in University campus and parking lot.

Figure 24 – Example of Background subtraction.

The brightness variation in the day, possibly due to solar occlusion by clouds, adverse weather conditions such as fog, rain, snow and wind, and the shadows of moving objects, Figure 25, are some of the many noise-inducing factors in the segmentation process. In these images, the points in the image classified as ghosts are updated in the reference images for the values of the color components define for those points. The remaining points are updated with a response filter to infinite impulse, in order to adapt the reference image to small variations of brightness [17].

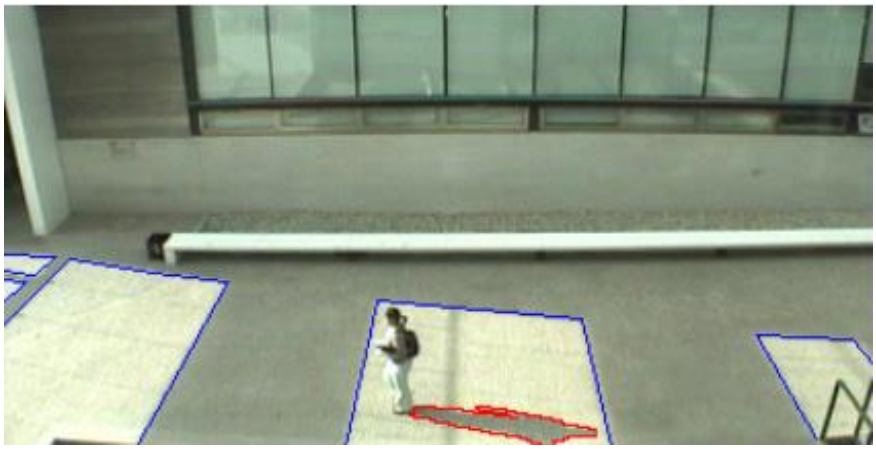


Figure 25 - Problem of shadows of moving objects [17].

3.3.2 Human tracking

In surveillance applications, tracking is the fundamental component. The human must first be tracked before recognition can begin. Kalman filters have been used extensively for tracking in many domains. The association of detections to the same object is based exclusively on motion. The motion of each track is estimated by a Kalman filter. The filter is used to predict the track location in each frame, and determine the likelihood of each detection being assigned to each track, as shown in Figure 26. Therefore, the assigned tracks are updated using the corresponding detections.

Figure 26 - Tracking of multiple objects in video surveillance.

Each track keeps count of the number of consecutive images, consequently if the count exceeds a specified thresholder, it will assume that the human left the scene and it will eliminate the track. By using the *initializeTracks* function it is possible to create an array of moving object in the video, so it will maintain the state of the tracked object until it leaves the scene.

For each tracked object, some information is collected, such as, the integer ID of the respective track, the number of frames since the track was first detected, the number of frames in which the track was first detected and the total number of consecutive frames in which the track was not detected.

To detect the human features in a sequence of image, it is necessary to convert the image into a binary image, as show in Figure 27, where pixels with the value 1 correspond to the foreground, and pixels with the value 0 correspond to the background scenario. It is performed a morphological operation on the resulting binary mask to remove the noisy pixels.

Figure 27 - Example of images converting to a binary image.

It was observed that the recognition and proper location of the human is not always possible in all frames, mostly because of the occlusion. However, the next step, is the assigning of the object detections in the current frame to existing tracks is done by minimizing cost. The cost is defined as the negative log-likelihood of a detection corresponding to a track.

3.3.3 Kalman filter

Subsequently, it was introduced into the system a Kalman filter that predicts where the human will be in the next frame considering the previous states of the object.

The Kalman filter operates in a two-step process, the prediction step and the updating step. In the prediction step the filter bases its output to the data obtained in time earlier using the system model which is defined initially and to MxN matrix that model the noise expected from the process, where M is the number of tracks, and N is the number of detections.

Solving the assignment problem represented by the cost matrix using the *assignDetectionsToTracks* function. The function takes the cost matrix and the cost of not assigning any detections to a track.

The value for the cost of not assigning a detection to a track depends on the range of values returned by the distance method of the vision.KalmanFilter. This value must be tuned experimentally. Setting it too low increases the likelihood of creating a new track, and may result in track fragmentation. Setting it too high may result in a single track corresponding to a series of separate moving objects.

3.3.4 Binary Image

After this process, has been accomplished it becomes necessary to convert the binary image into their skeleton, to subsequently detect the human features. This step is important, because if the skeleton of the human is not consistent with the image of the human, the features detection will not be very accurate, posture extraction. For this step, it was used the function:

$$BW2 = bwmorph(BW, operation, n); \quad (8)$$

This function applies a specific morphological operation to a binary image *BW*. In Figure 28, it shows an overall representation of the developed algorithm.

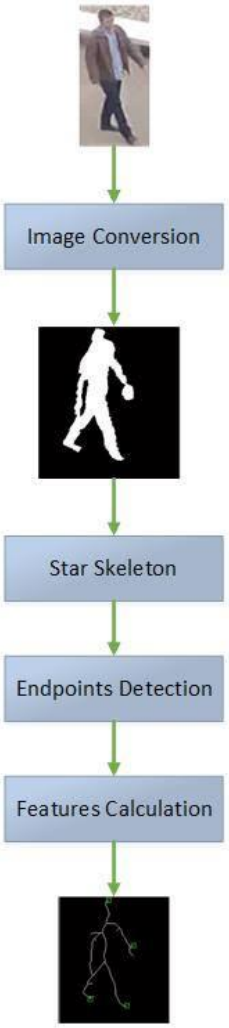


Figure 28 – Algorithm for human features detection.

3.3.5 Features extraction

As mentioned above, the human features to detect are the physical features of the human body, as for example, the head, the two hands and the two feet. Therefore, the strategy adopted for this case, was to determine the endpoints in the image, by using the operation: *endpoints*.

After the endpoints are already identified, the next stage will be detecting the other points of interest. Therefore, in Figure 29, it is possible to analyze the features detection algorithm that was used in this dissertation.

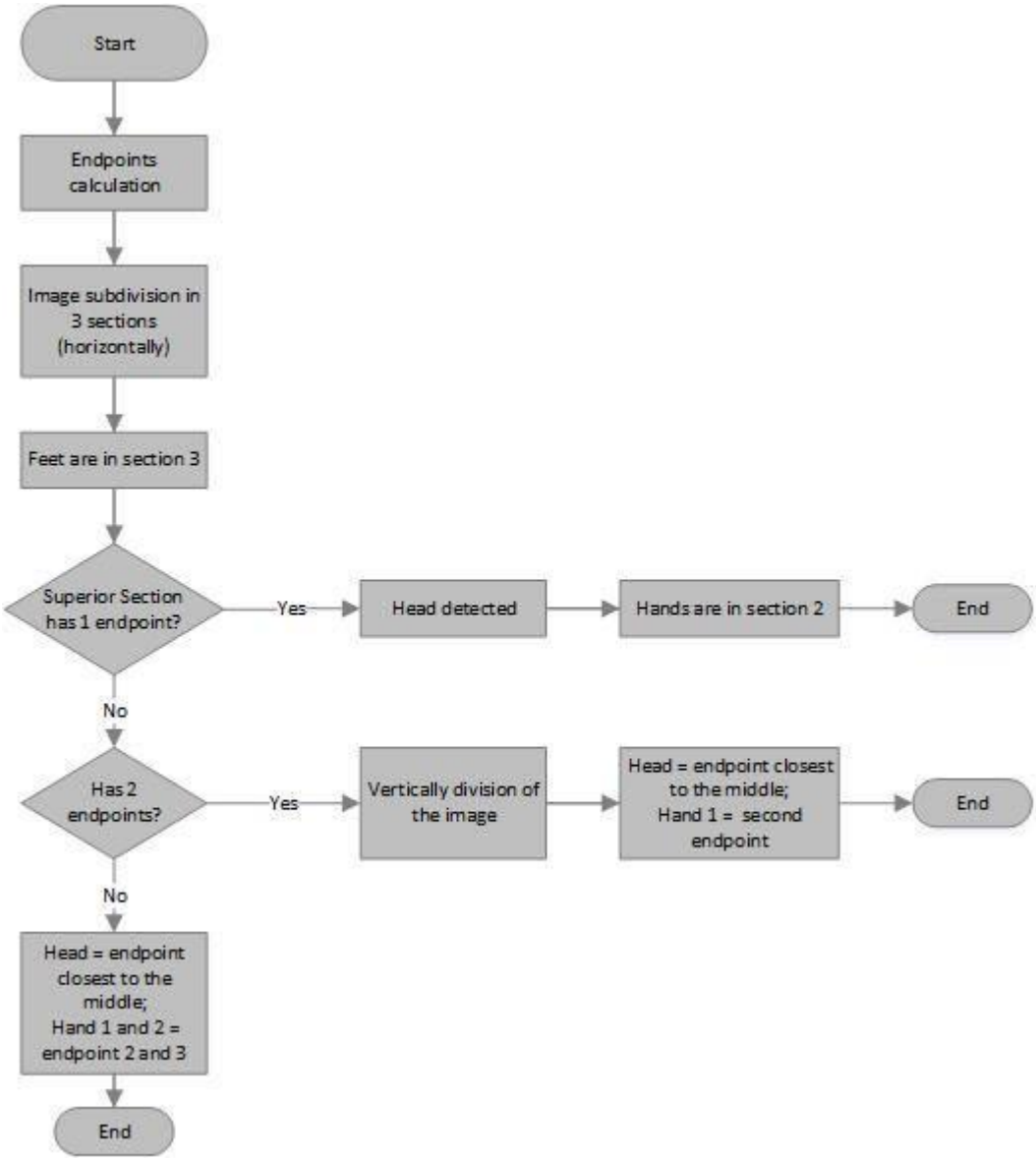


Figure 29 – Features detection algorithm.

4. RESULTS

In this chapter, it is described the results of the tests and the performance analysis of the algorithm for detecting the human features in a video surveillance.

4.1 Analysis of results

The posture of the human is considered important for classification of human actions.

These calculations of the human features must be run as fast as possible to ensure a maximum performance by the system. To analyze the developed algorithm, it was used 67 images captured from two cameras, HIK Vision and IR Network, in an indoor and outdoor environment. Also, all experiments were conducted using a common Asus laptop with 2.59 GHz Intel Core i7 processor.

For the evaluation of the algorithm it has been tested the number of features detected in the images. These images were previously captured by the object segmentation. To compare the results of the analysis, it was considered the number of features detected in each image, as shown in Table 4. With the results presented in Table 4, it can notice that the best results can detected five physical features (i.e. head, two hands and two feet). In six of the 67 images tested, it was possible to detect five physical features.

Table 4 – Analyze of the number of features detected in 67 images.

Number of features detected	**Number of images able to detect the feature**	**Image sample identification**
0	0	none
1	1	33;
2	15	13;21;23;29;30;34;35;36;53;56;61;62;65;66;67;
3	21	1;2;4;5;6;7;8;9;11;14;15;20;31;42;43;47;51;58;60;63;64;
4	24	3;12;16;17;18;19;22;26;27;28;32;37;40;41;44;46;48;49;50;52;54;55;57;59;
5	6	10; 24; 25; 38; 39; 45;

The calculation of the success of this algorithm was not compared with other systems, such as *Pfinder*, Microsoft Kinect, Star Skeleton and W^4. From the literature studied the results presented for these systems address to the next stage, the classification of human action.

In order to present the results in a more explicit way, it two different graphics were used. In each of them, the average results (detection of three to four features) and the worse results (detection of two to zero features) respectively.

As shown in Figure 30, there was cases that only just three and four human features were detected by the algorithm. In some images, it could be notice that the angle present by the human could not be perfectly recognized by the algorithm, being some human body parts hidden from the camera.

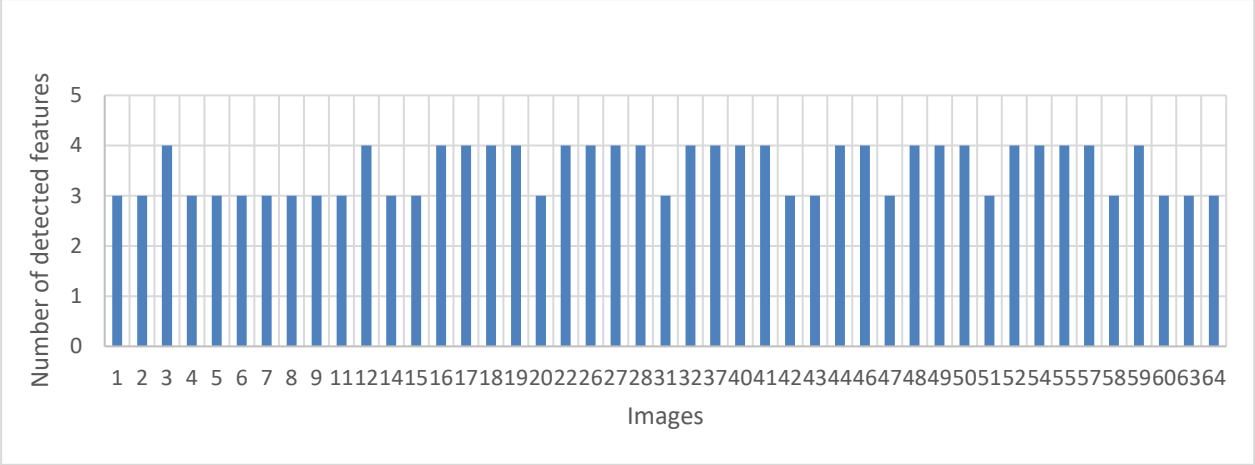


Figure 30 – Average results between 3 and 4 features detected.

In Figure 31, it is possible to visualize an example of four features detected in an image.

Figure 31 – Test example of four features detected.

Lastly, the worst results obtained were between two and one features detected. These results are due to the inability of the algorithm to distinguish certain objects, such as umbrellas or backpacks from the human body (Figure 32).

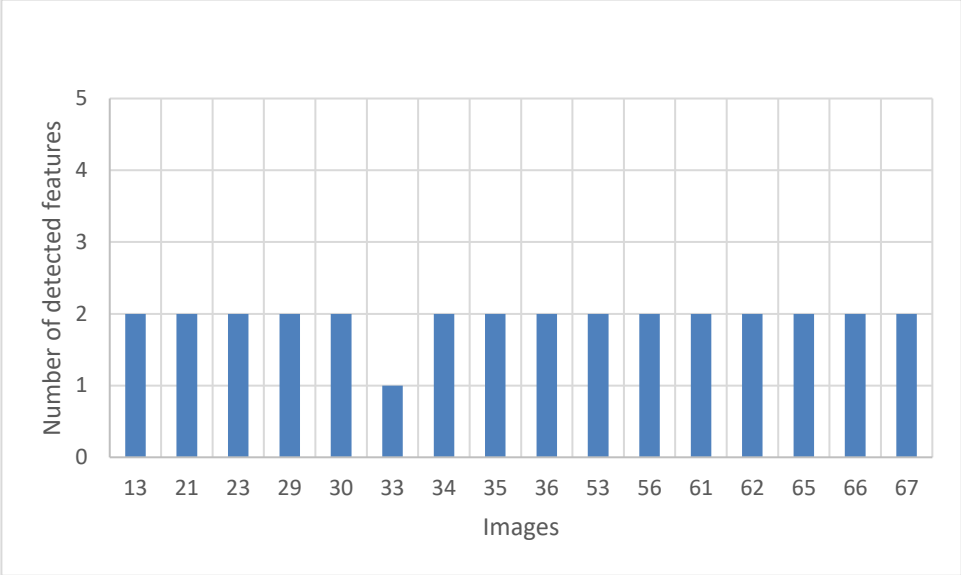


Figure 32 – The worse cases for detecting human features.

As shown in Figure 33, the average time obtained was 5910µs with a standard deviation of 5650µs.

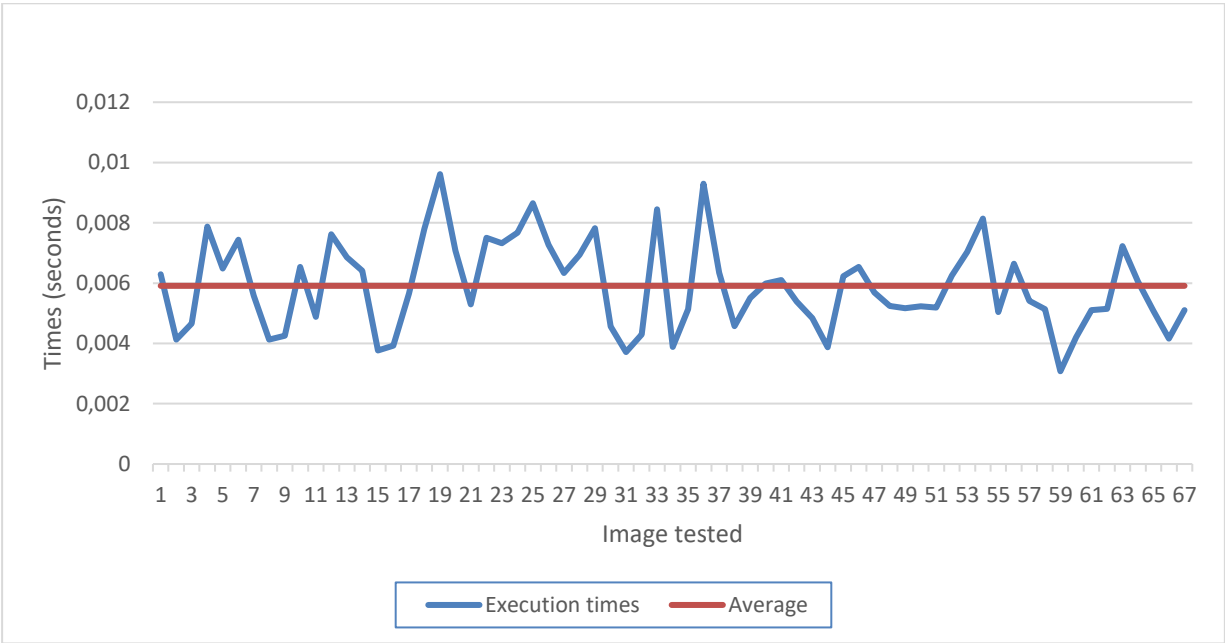


Figure 33 - Features calculation execution time.

4.2 Results analysis conclusions

A digital video consists of a sequence of static digital images. From the obtained results from 67 tested images, the success rate achieved was 8.96% (detection of five human features) and the average rate was 68.65% (detection of three to four features). Through these results it is possible to conclude that the algorithm was not 100% efficient to detect the five human features. Although, if it had been used a classification model it could predict the human action. In other words, by predicting five features in the first image, in the second image it could predict the position of all the features even when the algorithm just had detected three or four features before.

As stated before, such results could derivate from the conditions the images were recorded (i.e. weather and camera angle). Therefore, in the video there were many cases of people with umbrellas, heavy coats, hands in the pocket, among others, as shown in Figure 34. Nonetheless, for the creation of an effective video vigilance system the algorithm needs to be further explored toward solving these problems.

Figure 34 – Example of difficulties for features detection in images.

As shown in Table 5, all the human action recognition systems studied on Section 2.5, *Pfinder*, W^4, Microsoft Kinect and Star skeleton present some limitations. In other words, it is not completely effective for human action recognition.

As mentioned before, the main objective of this work is to present an algorithm that may contribute for the development of a human action recognition system that could be more effective than the existing systems. Consequently, it is not possible to compare correctly an algorithm, that is just a part of a system with an existing system.

Table 5 – Advantages and disadvantages of the systems.

System	**Limitations**	**Advantages**
Pfinder [3]	Recognition of a multiple objects; Multi-background; Dynamic scenes; Huge changes in lightning;	Few changes in the lightning; Real-time system; Occlusion;
W^4 [4]	Shadows; Low efficiency in a real scenario; Low light level environments; Recognition for upright people; Detection in long distances;	Multiple objects detection; Distinction between an object from the human; Occlusion; Low level processing;
Microsoft Kinect [43]	Detection in long distances;	Efficiency;
Star Skeleton [49]	Multipe objects detection; Shadows;	Can be used in animals and vehicles; Computationally cheap; It relies on no priori human model;

5. CONCLUSIONS

This last chapter presents the conclusions reached at the end of the dissertation, as well as the proposals for future work improvements.

5.1 Final conclusions

Automatic human action recognition systems have been widely researched within the computer vision and image processing communities. It is an active research field, due to its importance in a wide range of application, such as intelligent surveillance, visual surveillance, crowd behavior analyses, tracking of an individual in crowed scene, security breach, among others.

In a video sequences of a moving person, acquired with colored video, monocular and fixed camera systems, can be used to recover and detect human actions from a long distance.

The main goal of this dissertation is to present an algorithm, in order to contribute for the development of a human action recognition system that could be more effective than the existing systems. In this work, it was shown that after the detection of the human in an image it is possible to find five physical features (head, two hands and two feet). After, analyzing the results of this algorithm, it presents an average time of 5910 μs with a standard deviation of 5650 μs. Also, it presents a success rate of 8.96% (five features detected) and the average rate of 68.65% (three and four features detected). However, in this algorithm the results were not totally successful, to detect human features, mostly because of the weather conditions, in the time that the images were captured.

A digital video consists of sequences of static digital images called frames. Despite the fact, that the algorithm could not detected 100% the human features, it does not mean that in a sequence of images, the classification model could not predict properly the human action. For example, in a sequence of images, the first image detected five physical features (head, hands and feet) and the next one just four (head, one hand and feet), even though the classification model could estimate the position of the hand.

Even though the approaches based on skeletal models tend to perform with high recognition rates, extracting skeletal information from 2D videos is generally very difficult, primarily because of occlusions and distortion of human shape due to umbrellas, heavy coats, hands in the pocket.

There are some systems, such as *Pfinder*, W^4, Microsoft Kinect and Star skeleton. However, all these systems present some limitations, as for example, problems with shadows, lightning, and long distance, among others, not being fully effective for human action recognition.

Although, it presents some problems, it can be improved by adding another algorithm. Also, to achieve the best classification performance, the dimensionality of the feature vector should be as small as possible, keeping only the most salient and complementary features. In addition, keeping the dimensionality small could reduce the computational cost such that the recognition algorithms can be implemented and run on lightweight wearable devices such as mobile phones.

5.2 Future work

In this section are presented a few suggestions to complement this work.

This dissertation does not point to a single path for the recognition of abnormal behaviors observed by video surveillance, and there are many other ways that could lead to better results. In future works, the human features detection can be used for calculation or estimation of some motion body parameters, such as, the statistic velocities or accelerations, the angles of the feet and hands, trajectory of the human, among others, becoming possible to represent diverse human actions, for example, climbing, running and jumping. In other words, it will be necessary to implement a model to classify and to use this dataset (collected in University of Minho) for detect the human actions.

Another possibility is using the recovered shape and motion parameters to create a realistic animation or to produce automatically, with minimal human intervention, realistic animation models given a set of video sequences.

Nowadays, the algorithms are applied just on sequenced images of full bodies in motion. However, it is necessary to investigate the possibilities of having in the model guide a tracking process. If a point on the body's surface vanishes due to occlusion we can employ the model to predict where and when it will appear again. The likelihood of tracking errors can be reduced by using a more complex motion model, such as constant acceleration, or by using multiple Kalman filters for every object. Also, other cues can be incorporated to associate detections over time (i.e. size, shape, and color).

REFERENCES

[1] J. Ponce and D. Forsyth, *Computer vision: a modern approach*. 2012.

[2] D. Weinland and D. Weinland, "A Survey of Vision-Based Methods for Action Representation , Segmentation and Recognition," 2010.

[3] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, 1997.

[4] L. S. Davis, I. Haritaoglu, D. Harwood, and L. S. Davis, "W4 : Who ? When ? Where ? What ? A Real Time System for Detecting and Tracking People A Real Time System for Detecting and Tracking People W 4 : Who ? When ? Where ? What ?," April, 1998.

[5] H. Fujiyoshi and A. J. Lipton, "Real-time human motion analysis by image skeletonization," *Proc. Fourth IEEE Work. Appl. Comput. Vision. WACV'98*, pp. 15–21, 1998.

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

[7] D. Zhang and G. Lu, "Segmentation of moving objects in image sequence: A review," *Circuits, Syst. Signal Process.*, vol. 20, no. 2, pp. 143–183, 2001.

[8] P. Smith, D. B. Reid, C. Environment, L. Palo, P. Alto, and P. L. Smith, "Full-Text," vol. 20, no. 1, pp. 62–66, 1979.

[9] B. Horn and B. Schunck, "'Determining optical flow': A retrospective," *Artif. Intell.*, vol. 59, no. 1–2, pp. 81–87, 1993.

[10] J. Klappstein, T. Vaudrey, C. Rabe, A. Wedel, and R. Klette, "Moving object segmentation using optical flow and depth information," *Pacific Rim Symp. Image Video Technol.*, pp. 611–623, 2009.

[11] M. Narayana, A. Hanson, and E. Learned-Miller, "Coherent Motion Segmentation in Moving Camera Videos Using Optical Flow Orientations," *Comput. Vis. (ICCV), 2013 IEEE Int. Conf.*, pp. 1577–1584, 2013.

[12] A. J. Lipton, H. Fujiyoshi, and R. S. Patil, "Moving target classification and tracking from real-time video," *Proc. Fourth IEEE Work. Appl. Comput. Vision. WACV'98 (Cat. No.98EX201)*, vol. 98, no. 2, pp. 8–14, 1998.

[13] "Real-time Scene Stabilization and Mosaic Construction.pdf." .

[14] S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard people: a parameterized model of articulated image motion," *Proc. Second Int. Conf. Autom. Face Gesture Recognit.*, vol. 1121, pp. 1–7, 1996.

[15] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," *Proceedings. 1998 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (Cat. No.98CB36231)*, pp. 8–15, 1998.

[16] B. Galvin, B. McCane, K. Novins, D. Mason, and S. Mills, "Recovering Motion Fields: An Evaluation of Eight Optical Flow Algorithms," *Procedings Br. Mach. Vis. Conf. 1998*, p. 20.1-20.10, 1998.

[17] Q. D. S. Vídeo-vigilância, "Universidade do Minho Escola de Engenharia Previsão e Identificação de Eventos de," 2008.

[18] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, 1994.

[19] M. Piccardi, "Background subtraction techniques: a review," *2004 IEEE Int. Conf. Syst. Man Cybern. (IEEE Cat. No.04CH37583)*, vol. 4, pp. 3099–3104, 2004.

[20] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *Proc. 1999 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Cat No PR00149*, vol. 2, no. c, pp. 246–252, 1999.

[21] T. Bouwmans, F. El Baf, and B. Vachon, "Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey," *Recent Patents Comput. Sci.*, vol. 1, no. 3, pp. 219–237, 2008.

[22] D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organization of three-

dimensional shapes.," *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain)*, vol. 200, no. 1140. pp. 269–294, 1978.
[23] F. Lv and R. Nevatia, "Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching," *Comput. Vis. Pattern Recognition, 1992. Proc. CVPR '92., 1992 IEEE Comput. Soc. Conf.*, pp. 1–8, 2007.
[24] P. W. and Simon, "Too Big to Ignore : The Business Case for Big Data," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.
[25] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 53, no. 9. 2013.
[26] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6313 LNCS, no. PART 3, pp. 635–648, 2010.
[27] K. P. Murphy, "Machine Learning: A Probabilistic Perspective," *MIT Press*, p. 25, 2012.
[28] A. C. Lorena and A. C. P. L. F. de Carvalho, "Uma Introdução às Support Vector Machines," *Rev. Informática Teórica e Apl.*, vol. 14, no. 2, pp. 43–67, 2007.
[29] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," *Proceedings CVPR '92., 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 379–385, 1992.
[30] P. Afsar, "An Integrated System for Human Action Recognition from Video using Hidden Markov Model."
[31] A. Jain, J. Tompson, Y. LeCun, and C. Bregler, "MoDeep: A Deep Learning Framework Using Motion Features for Human Pose Estimation," *Comput. Vis. – ACCV 2014 SE - 21*, vol. 9004, pp. 302–315, 2015.
[32] F. Lerasle, G. Rives, and M. Dhome, "Human Body Limbs Tracking by Multi-ocular Vision," in *In International Conference on Pattern Recognition*, 1998.
[33] D. Meyer, J. Denzler, and H. Niemann, "Model based extraction of articulated objects in image sequences\nfor gait analysis," *Proc. Int. Conf. Image Process.*, vol. 3, no. Informatik 5, pp. 78–81, 1997.
[34] O. Munkelt, C. Ridder, D. Hansel, and W. Hafner, "A Model Driven 3D Image Interpretation System Applied to Person Detection in Video Images," in *In International Conference on Pattern Recognition*, 1998.
[35] C. Yaniz, J. Rocha, and F. Perales, "3D Region Graph for Reconstruction of Human Motion." 1998.
[36] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1–2, pp. 1–39, 2010.
[37] R. Plankers, P. Fua, and N. D'Apuzzo, "Automated body modeling from video sequences," *Model. People, 1999. Proceedings. IEEE Int. Work.*, pp. 45–52, 1999.
[38] C. R. Wren and A. P. Pentland, "Dynaman : A Recursive Model of Human Motion," *J. Image Vis. Comput. Spec. issue Face Gesture Recognit.*, 1998.
[39] H. J. Lee and C. Zen, "Determination of 3D Human-Body Postures From a Single View," *Comput. Vis. Graph. Image Process.*, vol. 30, no. 2, pp. 148–168, 1985.
[40] C. R. Wren, F. Sparacino, A. J. Azarbayejani, T. J. Darrell, T. E. Starner, A. Kotani, C. M. Chao, M. Hlavac, K. B. Russell, and A. P. Pentland, "Perceptive Spaces for Performance and Entertainment : Untethered Interaction using Computer Vision and Audition," vol. 11, no. 372, 1997.
[41] T. Starner and a Pentland, "Real-time American Sign Language recognition from video using hidden Markov models," *Comput. Vision, 1995. Proceedings., Int. Symp.*, pp. 265–270, 1995.
[42] A. De and N. Do, "Reconhecimento de Objetos Baseado em Visão Artificial," 2015.
[43] S. Prabhu, B. E. Extc, J. Kumar, B. B. E. Extc, A. Dabhi, and P. Shetty, "Real Time Skeleton Tracking based Human Recognition System using Kinect and Arduino," *Int. J. Comput. Appl.*, pp. 975–8887.

[44] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, “Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition,” *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 24–38, 2014.

[45] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1290–1297, 2012.

[46] J. Webb and J. Ashley, *Beginning Kinect Programming with the Microsoft Kinect SDK*. 2012.

[47] M. R. Andersen, T. Jensen, P. Lisouski, A. K. Mortensen, M. K. Hansen, T. Gregersen, and P. Ahrendt, “Kinect Depth Sensor Evaluation for Computer Vision Applications,” *Electr. Comput. Eng.*, vol. Technical, p. 37, 2012.

[48] C. Eisler, “7750,” *Kinect for Windows Team*, 2012. [Online]. Available: http://blogs.msdn.com/b/kinectforwindows/archive/2012/01/20/. [Accessed: 30-Oct-2016].

[49] D. Singh, A. K. Yadav, and V. Kumar, “Human Activity Tracking using Star Skeleton and Activity Recognition using HM M ’ s and Neural Network,” vol. 4, no. 5, pp. 1–5, 2014.

[50] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3D skeletons as points in a lie group,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 588–595, 2014.

[51] “Recognition and Classification of Human of Human Behavior in Intelligent Surveillance Systems.pdf.crdownload.” .

[52] K. Aziz, D. Merad, B. Fertil, and N. Thome, “Pedestrian head detection and tracking using skeleton graph for people counting in crowded environments,” *Conf. Mach. Vis. Appl.*, no. January 2011, pp. 516–519, 2011.

[53] H. Mo, J. Leou, and C. Lin, “Human Behavior Analysis Using Multiple 2D Features and Multicategory Support Vector Machine,” *Computer (Long. Beach. Calif).*, pp. 0–3, 2009.

[54] J. Pradilla, “Matlab 7,” p. 1, 2005.

[55] C. Mathworks and A. H. Drive, “Computer Vision System Toolbox ™ Getting Started Guide R 2015 a,” 2015.