Faculdade de Engenharia da Universidade do Porto

Extraction and Identification of Extremities for Silhouette Representation on Individual Actions

Miguel Angelo Oliveira Couto

VERSÃO FINAL

Dissertação realizada no âmbito do
Mestrado Integrado em Engenharia Electrotécnica e de Computadores
Major Telecomunicações

Orientador: Daniel Cardoso de Moura, PhD
Co-orientador: Eduardo Marques, Eng.

Julho 2014

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

MIEEC - MESTRADO INTEGRADO EM ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES | **2013/2014**

A Dissertação intitulada

"Extraction and Identification of Extremities for Silhouette Representation on Individual Actions"

foi aprovada em provas realizadas em 18-07-2014

o júri

Presidente Professor Doutor Jaime dos Santos Cardoso
Professor Auxiliar do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

Professora Doutora Ana Maria Perfeito Tomé
Professora Associada do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro

Doutor Daniel Cardoso de Moura
Investigador Auxiliar do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

O autor declara que a presente dissertação (ou relatório de projeto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extratos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são corretamente citados.

Autor - Miguel Ângelo Oliveira Couto

Faculdade de Engenharia da Universidade do Porto

Resumo

A representação de silhuetas Humanas em vídeo pode ser vista como um passo importante em direção à classificação de ações. Esta tese aponta para a exploração de um método para descrever o corpo Humano em vídeo, nomeadamente através da deteção de 5 pontos de extremidade nos contornos da silhueta, mais propriamente na cabeça, mãos e pés. Adicionalmente, 4 ângulos que relacionam os pontos de extremidade das mãos e cabeça e dos pés e centroide dos contornos são propostos como um descritor que fornece informação anatómica que pode apresentar-se como útil para trabalhos futuros de reconhecimento de ação. Por outro lado, um método baseado no Chain Code de Freeman, que tem como output um conjunto diferente de pontos de extremidade, é proposto ser usado como um pós-processamento de maneira a potencialmente melhorar a deteção e identificação dos 5 pontos de extremidade.

O algoritmo foi testado no dataset "Weizmann" que contém sequências de vídeo de indivíduos efetuando 10 ações diferentes. Os 5 pontos de extremidade foram identificados manualmente utilizando o Matlab em 30 vídeos de 3 sujeitos diferentes levando a cabo cada uma das referidas ações, de maneira a providenciar os pontos de extremidade de ground-truth para este estudo. Os resultados foram de seguida avaliados através da comparação dos pontos de extremidade detetados e os pontos de ground-truth correspondentes. A distância Euclidiana entre os pontos detetados e os pontos de ground-truth foi medida como erro que foi normalizado como um fator da altura do sujeito (cujo cálculo foi baseado nos pontos de ground-truth da cabeça e dos pés numa posição ereta).

Os resultados mostram que independentemente da ação, uma precisão de ~90% foi atingida para os pontos de extremidade da cabeça e pés. Contudo, não foi possível atingir este nível de resultados para as mãos, aparte de alguns tipos de ação. As ações em que a deteção das mãos atingiu melhores resultados incluíam um movimento das mesmas que implicava que se posicionassem afastadas do torso. O facto de a sua forma estar oculta no interior da silhueta na maioria dos vídeos analisados, revelou ser uma barreira para a correta deteção dos pontos de extremidade correspondentes.

A subtração de background revelou ter algum impacto negativo na aplicação do algoritmo, especialmente em casos onde os sujeitos não se evidenciavam claramente do plano de fundo. Uma diferença máxima de ~15% nas medidas de precisão obtidas com a utilização das máscaras de ground-truth e com subtração de background mostra que uma segmentação de foreground mais eficiente poderia resultar numa melhor eficácia do algoritmo. Adicionalmente, através de análise visual dos padrões gerados pelo descritor angular proposto nesta tese mostra resultados promissores em direção a classificação de ações.

Globalmente, o pós-processamento não melhorou os resultados atingidos com o algoritmo proposto. A secção de Trabalho futuro inclui a proposta de uso de um critério mais complexo a ser aplicado ao pós-processamento, que leva em conta informação anatómica do corpo Humano.

Abstract

Human silhouette representation in video may be used as an important step towards the classification of actions. This thesis aims to explore a mean of describing the human body on a video sequence, namely by detecting and identifying 5 extremity points amongst the silhouette contours in the head, hands and feet. Additionally, four angles relating the hands and head and the centroid of contours and the feet are proposed as a descriptor that provides further anatomical data which may present itself useful for future action recognition works. On the other hand, a Freeman Chain Code based method, which outputs a different set of feature points, is proposed to be used as a post-processing in order to potentially improve the detection and identification of the proposed 5 extremity points.

The algorithm was tested in the "Weizmann" dataset containing video sequences of single individuals performing 10 different actions. The 5 feature points were manually identified using Matlab in 30 videos of 3 different subjects conducting each of the actions in order to provide the ground-truth feature points of this study. The results were then evaluated by comparing the detected feature points to their reference position. The Euclidean distance between the detected feature point and its manually identified position was measured as an error which was normalized by a factor of the subject's height (whose calculation was based on the head and feet reference points).

Results show that independently from the action, precision rates of ~90% were achieved for the feet and head. However, it was not possible to achieve such results for the hands, apart from some specific actions. The actions in which the detection of the hands feature points performed better included a motion of the hands in which they were well set apart from the torso. The fact that their shape was occluded on the inside of the silhouette in most analyzed videos revealed to be a constraint for their feature point detection.

The background subtraction revealed itself to have some negative impact on the application of the algorithm, especially in cases where the subjects did not clearly stand out from the image background. A maximum difference of ~15% on the precision rates using background subtraction and

the ground-truth masks shows that a more effective foreground segmentation could indeed result in a better performance of the algorithm. Additionally, visual analysis of the patterns generated by the angular descriptor proposed in this thesis shows promising results towards actions classification.

Overall, the post-processing did not improve the results achieved with the proposed algorithm. Future work includes using a more complex criteria to apply the post-processing that takes into account anatomic information of the Human body.

Agradecimentos

O meu sincero obrigado aos meus orientadores. Daniel Moura e Eduardo Marques, que sempre me apoiaram e orientaram ao longo do trabalho que desenvolvi para esta tese. Obrigado por, juntamente com o Pedro Cunha, sempre terem estado dispostos a ajudar-me a evoluir e aprender a desenvolver trabalho científico.

Obrigado aos que me apoiaram nestes 5 anos, aos meus colegas e amigos com quem segui este passo da minha vida que culmina com esta tese.

À minha família, que esteve lá nos momentos difíceis. Aos meus avós por tudo o que representaram ao longo de toda a minha vida. À minha mãe.

Miguel Couto

"Live as if you were to die tomorrow. Learn as if you were to live forever."

Mahatma Ghandi

Contents

List of Figures

List of Tables

Abbreviations and Symbols

BS	Background Subtraction
CS	Contour Signature
CV	Computer Vision
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DT	Distance Transform
EDS	Equidistant Sampling
EPS	Extreme Point Sampling
FD	Fourier Descriptor
FP7	European Union Seventh Framework Programme for Research
HoG	Histogram of Oriented Gradients
HU	Hu Moments
MOG	Mixture Of Gaussians
OM	Occupancy Map
PF	Poisson Features
SC	Shape Contexts
SM	Saliency Map
SVM	Support Vector Machines
SS	Start Skeleton
VSS	Variable Star Skeleton

Chapter 1

Introduction

It is no news the enthrallment Humanity has devoted to the ability to see by a non-living entity. For instance, it is clear in the Greek mythology one of the first allusions to artificial vision, where a giant named *Talos* was created and presented as a gift by the ancient god Hephaestus to Minos, the King of Crete. He was supposed to be a law agent of the island, responsible for making sure no harm came to the inhabitants as well as applying the law of the land [1].

Even though probably men once thought artificial vision only had room in mythology or fiction, it is a reality nowadays. In fact Computer Vision (CV) has been a quite increasingly growing area of research and has also been dividing itself into many different sub-areas. As a vast part of the human brain is directly or indirectly related to vision tasks and being vision itself the most resource consuming human sense [2], it should be no surprise that finding ways to ease the image processing and interpretation job of the brain has become the main purpose of many researchers.

In order to analyze Human behavior in video, a main task that can provide valuable input for such analysis is the silhouette representation. It is in this context that this thesis is positioned. It aims to propose a way of representing a Human silhouette through the extraction of extremities while providing specific information about it. The contributions of this work can be summarized on the following topics:

- Proposal of an algorithm capable of extracting and identifying feature points of 5 extremities of the Human silhouette: head, both feet and both hands;
- Proposal of 4 angles which provide anatomic information about the pose of the subject;
- Proposal of a post-processing based on a Freeman chain code algorithm that can possibly improve the feature points location.

The current chapter is divided into three sections. In Section 1.1 the context in which the project was proposed is explained. In Section 1.2 the final purpose of the upcoming work is detailed. And ultimately, a reader's guide is provided in Section 1.3, where the reader is led through the organization of the document.

1.1 - Motivation

This thesis is developed under the *FP7* funded project Future Cities. The main purpose of the latter is basically making the Portuguese city of Porto "an urban-scale living lab" [3], *i.e.* facilitating the city technological means capable of gathering information in real time and to provide interconnectivity in innovative ways, for instance by taking advantage of Android smartphones to get vital signs (from volunteer users) or by building a Vehicular Ad-hoc Network over harbor trucks and nearly 500 buses. In this case, the technology being used for communication between vehicles is either 802.11p, WiFi or 3G which shall prepare them for also interact with the main infrastructure, taking into consideration the already known location of public Wifi hotspots and Road Side Units thus having the ability to send information to the cloud, whenever there are conditions to do so.

Another pillar of development of the project consists in building a platform of monitoring devices that are able to collect data from mainly video and audio sources. At this point it is hereby being planned the implementation of over 50 cameras connected to small single-board computers, namely the Raspberry Pi. The long-term expectation is to provide automated action recognition therefore identifying specific situations (*e.g.*, possible robberies, aggressions). Thereby, a major aspect of computerized action classification is the silhouette extraction from the individual, which is precisely where the outcome of this thesis stands.

Nevertheless, for the purposes of this work, the conducted experiments involved a more simplistic environment that involved one single subject In each video and a non-complex background *i.e.* a background that did not include many details, hence facilitating a smoother foreground subtraction.

1.2 - Objectives

The silhouette representation plays an important part in a system that aims to characterize the behavior of a subject. In that sight, the major goal of this thesis is to propose a way of representing a Human silhouette providing relevant information about it, such as the location of certain points of interest or alternative measures relating them. The main objectives of this thesis are:

- Propose a method capable of extracting and identifying key anatomical feature points, potentially useful to characterize the pose of the subject;
- Propose a descriptor based on these feature points, which may be potentially useful to describe the action.

Identifying the body parts associated with each feature point is a step forward from merely extracting them from a Human silhouette. By successfully performing the identification, more discriminative descriptors may be proposed based on the identified feature points. Consequently, more discriminative descriptors may potentially lead to a more effective action classification.

Familiarization of the Future Cities Project framework was an integral part of the work, since it is crucial to be in tune with the investigation efforts of the team to the date. Additionally, a full review of the literature regarding methods to be used for the foreground segmentation and silhouette descriptors was conducted.

1.3 - Document Outline

In the current chapter an introductory overview of the thesis and the correspondent context is provided. In Chapter 2 a review of relevant subjects for this thesis is performed as a preparatory study for the upcoming work. Several descriptors are presented likewise a summary analysis of the skeletonization method of human silhouettes as an alternative representation of the latter.

In Chapter 3, the proposed algorithm is presented and positioned taking in consideration previous research. A post-processing step is also introduced as a mean to strengthen the overall results. In Chapter 4, the experiments conducted to test the algorithm are introduced and the results are presented and discussed taking in consideration several perspectives. The limitations of the algorithm are analyzed independently of external factors to an uninfluenced performance of the algorithm. For this purpose, the ground-truth masks are assumed to provide an accurate foreground segmentation which would be a requirement for an uninfluenced execution of the algorithm. The impact of background subtraction is then evaluated on the performance of the algorithm, the impact of the proposed post-processing is also evaluated and the proposed four angles are discussed as a silhouette representation feature.

Finally, the conclusions taken from the experiments are addressed in Chapter 5 and a link to a future research path is proposed.

Chapter 2

Literature Review

As conceptualized in [4], Computer Vision (CV) can be categorized as the following stages: Detection (presence of a given item in the scene), Localization (locate the particular item) and Recognition (locate all the given items in the scene) and Understanding (take into consideration the environment and understand the surrounding context). Within each of these stages, further levels of tasks can also be discriminated. All these aspects ought to be taken into consideration whereby a careful planning of the problem in the correspondent context. In this case, the primary objective is to perform Human shape representation, therefore it can be seen as a grounding work for the Understanding stage. It is as well a baseline work for further recognition processes. In order to understand an image, the first thing to do ought to be identifying the subjects and objects of interest within it. Appliances of this process can be found in several fields, which include video surveillance [5], [6], human tracking [7], [8], human-computer interaction scenarios [9], human body analysis in sports [10], traffic control [11], just to mention a few.

As it is known, the human body permits a wide range of well defined poses which, allied to disparities in body dimensions and appearance as well as important environment conditions (*e.g.* lightning conditions and sudden variations, camera position and resolution), may difficult the correct discrimination of the silhouette due, for instance, to a higher probability of occlusion either because of some part of the body is hidden from the point of view of the camera or because the light is so poor that a person wearing dark colors may even not be differentiated from a dark background. The most widely used method for the purpose of differentiating the acting subjects of a given scene from the background scenario is surely Background Subtraction (BS) [12]–[15].

Given a video in which one wish to extract the contour of the silhouette of an object, namely a human, two major steps ought to be taken into consideration. Firstly, it is important to discriminate the actual silhouette, differentiate it from the remaining scene. If the target is moving, then one of the best approaches would be using BS since its performance is increased the more motion the foreground object to be detected is conducting. If it is not the case and the target is still, the difficulty of the task may increase. In such a situation, a possible approach would be applying a

Saliency Map (SM) based method in which the visual saliency[1] of the object, person is evaluated (in which case the color, gradient and other image parameters of the person would have to stand out from the scene for a successful contour extraction). Background Subtraction as well as Saliency Maps (SM) based approaches are addressed in Section 2.1.

Secondly, after detecting the silhouette contour of the human, it is important to process that information for further analysis. That can be accomplished by contour descriptors which are introduced in Section 2.2. The descriptors provide information about the silhouette contours they describe. Besides, they can be used to measure the similarity between two contours, thus proving their selves useful in object recognition, by comparing an extracted contour to the ones in a public dataset containing contours of similar objects [16].

In Section 2.3 Skeletonization is addressed, which is an approach used to distinguish the location of body extremities, taking as input an Human silhouette contour. It is presented in an action recognition scenario, which as previously mentioned, it is referred as a further investigation trajectory following this research work.

Finally, an overview of the chapter is provided in Section 2.4. The relevancy of each addressed topic is evaluated considering the purpose of investigation of the thesis.

2.1 - Silhouette segmentation

The term “image segmentation” refers to the obtainment of a more simplified representation of an image in something easier to process and analyze[17]. Henceforth, “silhouette segmentation” refers to the process of obtaining a binary representation of a given image where the foreground pixels belong to the silhouette. In this Section, two approaches to perform silhouette segmentation are presented. Specifically, two main paradigms are presented by each one of them:

- Silhouette segmentation based on detected movement by the humans on the scene;
- Silhouette segmentation based on visual saliency methods.

Both are evaluated and reviewed, their strengths and the scenarios each method ought to present better results are discussed.

1. Visual saliency is the perceptual quality that makes an object, person, or pixel stand out relative to its neighbors and thus capture our attention. Visual attention results both from fast, pre attentive, bottom-up visual saliency of the retinal input, as well as from slower, top-down memory and volition based processing that is task-dependent [18].

2.1.1 - Background Subtraction

BS provides a simple way of identifying moving objects in a static background. The method is especially effective in scenarios where the camera is stationary as well as a reasonably noise-free background (e.g., no tree leaves moving, no ocean waves nor any other kind of movement that may wrongly disassociate an object from the background) is considered. Performance is a major concern, and has been one of the aspects subject of further research and improvement, as recently accomplished by Yiran Shen *et al.* [12], who reports performance enhancements of 5 times compared to conventional BS algorithms, as well as significantly less memory and energy consumption.

The rationale common to most BS methods begins with pixel analysis, whenever a given set of pixels are detected in different frames in different positions and yet their resemblance lead to conclude that they belong to the same object, then they are categorized as belonging to a moving object *i.e.* to the foreground. For such sorting, it becomes necessary to ensure a way of generating the so called background model. Nevertheless, several difficulties may arise against the good performance of its calculation. For instance, the low camera resolution, the already mentioned background noise, as well as gradual and especially sudden illumination changes or even camera jitter. These factors generate false positives, and on the other hand false negatives can occur as well whenever the camouflage effect takes place (a moving object is circulating in an area of the background where the colors of both are very similar, thus being wrongly considered part of the background).

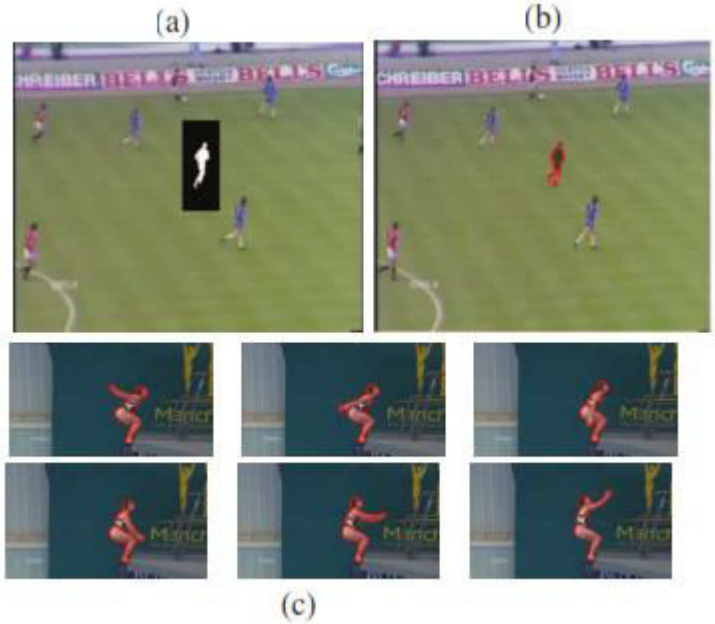


Figure 2.1: Background Subtraction applied to human silhouette contour extraction (Source: adapted from [19]) (a) BS applied in the detection window, (b) Extracted silhouette contour, (c) Human silhouette contour tracking in consecutive video frames

Taking in consideration the principles of BS, it is an approach whose efficiency and effectiveness largely depends on whether the foreground object is moving or not. It has been used in a recently published work by Ajoy *et al.* [19] in silhouette boundaries extraction as part of an integrated

human detection and contour tracking system. The approach consisted in using the HoG descriptor combined with a k-nearest-neighbor classifier as the human detection algorithm. Once the candidate window was delimitated, BS was applied exclusively in it, hence a binary indicator would pin the object pixels with ones and non-object pixels with zeros. The transition pixels between the background and the object silhouette were treated as part of the silhouette boundary. Finally, after connecting such pixels, the human contour would be extracted from the target frame. In Figure 2.1 are represented some results of the algorithm. In (a) is the result of the BS, after which the contour pixels are identified and connected as shown in the result in (b), the silhouette contour representation. In (c) a result of 6 consecutive video frames silhouette contour tracking is presented. The videos in which the algorithm was tested presented humans in constant movement, fact that explain the good results of BS for the silhouette contour extraction.

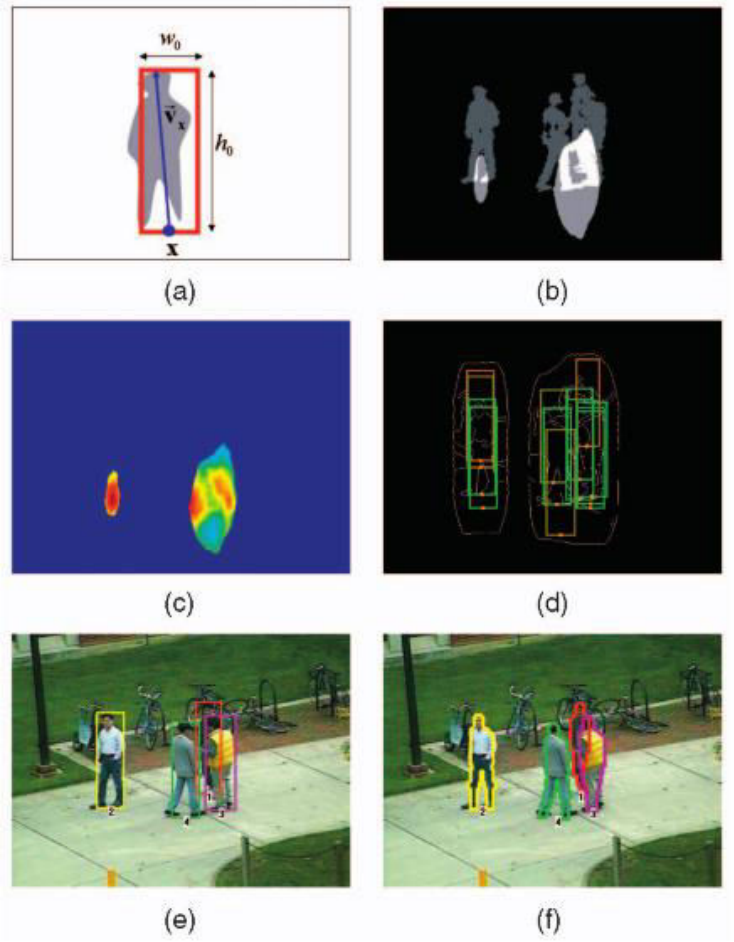


Figure 2.2: Background Subtraction as an optimization method in human silhouette Segmentation (Source: adapted from [20]) (a) Adaptive window used for detection, (b) Regions more likely to contain moving humans are represented with lighter colors, (c) Likelihood map, (d) set of detection windows (the green ones represent higher likelihoods whereas the red windows represent lower likelihoods, (e) Final detection results, (f) Final human silhouette segmentation

In order to make use of its strengths to add robustness and efficiency to the algorithm, Lin *et al.* [20] applied BS in a detection and subsequent silhouette contour extraction system thus optimizing the obtained results. A hierarchical template matching method where human detection is made by using global human shape models which are built by detecting several human body parts. As stated

by the authors, the used paradigm approaches human detection and silhouette contours extraction simultaneously. A descriptor very similar to HoG, gradient magnitude-weighted, is considered in order to extract the image features and also the silhouette contours. With the intention of improving the efficiency and robustness, BS is applied afterwards in order to narrow the human detection procedure. Additionally, the obtained region information by the foreground areas is combined with the shape information from the original image hence improving robustness. In Figure 2.2 the results of the silhouette contour extraction with BS are presented. Instead of performing a search in the whole image, the use of BS permits detecting the moving objects thus representing the regions in where most likely there would be a human. However, in complex scenes containing multiple humans and not ideal illumination, the authors reported a quite noisy background subtraction. Despite that fact, the extracted silhouette contours were accurate.

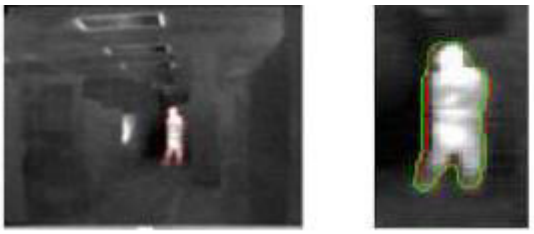


Figure 2.3: Extracted silhouette contours using BS followed by an active contour model (Source: adapted from [21])

BS has also been used for successfully extract human silhouettes in thermal infrared surveillance systems. In 2013, Tan *et al.* [21] proposed a two level set based active contour model[1] based in BS, which was combined with an edge detector[2] in order to improve segmentation accuracy however with a computational cost. In Figure 2.3 the resulting extracted silhouette contours are represented. Given the fact that infrared images have a very low quality, the results are quite satisfactory.

In a context where humans are in constant movement, which should occur in the expected general scenario (*i.e.* urban environments), BS is a prospective candidate method to perform silhouette segmentation. However, there exists the risk of a substantial amount of noise lead to a not so accurate performance.

1. The fundamental idea of an active contour model is to start with an dynamic contour around the object of interest, and then the contour moves toward its interior normal and stops on object boundaries. A formal mathematical presentation can be found in [21].

2. An edge detector is an algorithm that performs edge detection. Edge edtection is the name for a set of mathematical methods which aim at identifying points in a digital image at which the image brightness

changes sharply or, more formally, has discontinuities. The points at which image brightness changes sharply are typically organized into a set of curved line segments termed edges [22].

2.1.2 – Saliency Maps

The concept that inspire SM based algorithms is based on biological vision, namely visual attention which refers to one's ability to quickly differentiate the most important and evident information within a scene [23]. There are so many tiny details that even though our eyes capture, our brains cannot possible process them all, so they choose to process whatever stands out more and that is usually what is retained in our memories, unless of course we train our brains to be more attentive and observers. After all, as allegedly the famous artist Michelangelo said [24], "The Devil is in the details".

General SM methods are based in a low-level approach where contrast analysis of an image region, compared to the surroundings is considered. Normally factors like intensity,, color and orientation are evaluated. Such methods can be categorized in biologically based, purely computational or a combination of both. An example of the first is accomplished by Itti *et al.* [25] where by employing a Difference of Gaussians approach, the center surround contrast is calculated. Later Frintrop *et al.* [26] improved the method proposed by Itti by using integral images to gain computational efficiency and on the contrary of the latter, Frintrop used center-surround differences with square filters.

On the other hand, purely computational methods do not adopt biological vision principles. For instance, saliency has been estimated using several methodologies namely using center-surround feature distances [27], or by applying heuristic procedures on saliency measures [28]. Vasconcelos *et al.* maximize the mutual information between the feature distributions of center and surround features in an image and additionally Hou *et al.* [29] base their saliency detector in a frequency domain processing. As an exemplificative work that combines both approaches, Harel *et al.* [30] makes use of the previously mentioned Itti's method but the normalization is completed relying on a graph based approach.

Figure 2.4: Result of a SM approach followed by a binary map (Source: adapted from [18]) (a) original image, (b) Result of the salient region detector followed by a binary map

A quite relevant work worth mentioning is the one accomplished by Achanta *et al.* [18] which considers a frequency approach and combines image features such as color and luminance. Afterwards the saliency detection, a binary map is applied thus the pixels belonging the salient region are assigned with ones and the remaining pixels are assigned zeros. The authors claim very high efficiency as well as better performance taking in consideration a comparative study conducted relatively to other five state of art salient region detectors. In Figure 2.4, some results of the mentioned algorithm are presented. As seen, the output is similar to a noise-free BS based method. Of course the proposed images have quite distinguishable thus salient objects which explains the so good results. In real life situations, to be precise in urban scenarios that may not be the case. People might not stand out from the general scene, fact that may affect the performance of the algorithm.

Bearing in mind the same type of approaches used with BS in order to obtain the object silhouette contours, ideally considering an output of the SM algorithm the silhouette of the human, one can apply an active contour approach to perform the silhouette segmentation as accomplished by Tan *et al.* [21] after applying BS or as also successfully applied with simple BS [19] in which the transition pixels between the foreground and background are considered to be part of the silhouette contour.

SM approaches are advisable to situations where the target human is not moving. In a situation where the target human is completely still, BS is not effective thus SM may present itself as a valid alternative. However, its performance strongly depends on how distinguishable the human is in the scene. Roughly speaking, in a worst case scenario, if someone all dressed up in black in a very dark area and on top of that the person is completely stationary, it may very well be the case that none of the proposed approaches ought to be effective.

2.2 - Descriptors

Over the years, many algorithms have been conceived in order to delineate the shape of objects of a given image, more particularly Human shapes. In order to describe the shape contour information of Human silhouette there exists several algorithms and methods. The task of identifying the shape of the same subject or object over several sequent images is performed by matching algorithms. Such algorithms can be categorized into massive (Occupancy Maps (OM), Poisson Features (PF) and the first version of Hu Moments (HU)) which are based the entire content of the silhouette or contour based (Distance Transform (DT), Contour Signature (CS), Fourier Descriptor (FD), Shape Contexts (SC), Discrete Cosine Transform (DCT) and the second version of HU) that are strictly dependant on the silhouette contours.

Nevertheless, a different kind of descriptors are addressed in this topic, namely feature descriptors (Histogram of Oriented Gradients (HoG), Haar-Like Features and SC) which aim to represent characteristics other than the contours of a given shape.

2.1.1 - Fourier Descriptor

The rationale behind FD is to define n points over the boundaries of the silhouette $\{(x_0, y_0), \ldots, (x_{n-1}, y_{n-1})\}$ which are frequently sampled using equidistant sampling (EDS), being the distance along the silhouette between two consecutive points similar. Henceforth, they ought to be converted into complex coordinates $z_i, i \in [0, n-1]$ with $z_i = x_i + y_i\sqrt{-1}$ and subsequently transformed to its frequency domain by means of a DFT. After this step the Fourier coefficients are obtained, indicated by $\{f_0, \ldots, f_{n-1}\}$. The low index coefficients represent information regarding of general form of the shape whereas high index coefficients contain more detailed information of the shape. Being the first coefficient exclusively dependent of the shape, position invariance is achieved by considering it zero. Rotation invariance can be attained by diving the remaining coefficients by the second f_1, ergo after normalization one can consider $n-2$ unique coefficients since $f_0 = 0$ and $f_1 = 1$, which are given by (2.1).

$$FD = \frac{|f_i|}{|f_1|}, i \in [2, n-1] \tag{2.1}$$

Poppe and Poel [31] propose an approach useful to refine the sampling, it aims to retrieve points of extreme curvature so that they are prioritized amidst the sampling. Since points of extreme curvature (*e.g.* hands or head) contain more information, consequently points of minimum curvature (*e.g.* neck or torso) would contain less information. If not a sufficient number of coefficients and local noise are considered, there may be the case of several points of extreme curvature not being sampled. In order to deviate this restrain, it is proposed to set the first 200 FDs to zero to eliminate noise. Afterwards, the number n of extreme points is calculated. For a m obtained FDs, if $n < m$, the number of FDs are lowered until $n > m$ and $k - m$ additional points

such that the summed square distance between all remainder n pairs of the silhouette is minimized. This process is called Extreme Point Sampling (EPS). In Figure 2.5 it is clear to see the sampling improvement considering the maximization of extreme curvature points. The most evident are the point sampled with EPS in the head and the two on both feet which were not sampled with EDS.

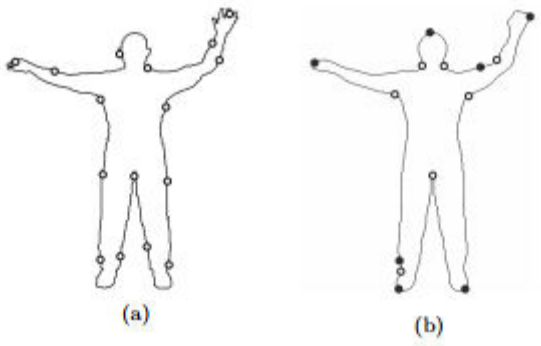


Figure 2.5: FD sampling using EDS and EPS (a) 16 EDS sampled points along the silhouette, (b) Reconstructed sampling with EPS using 84 Fourier coefficients. Points with extreme curvature are filled circles (Source: adapted from [31])

In order to measure the similarity of two given FD, $FD1$ and $FD2$, one can use the Euclidian distance between them, as presented in (2.2).

$$d = \sqrt{\sum_{i=0}^{n-3} |FD1_i - FD2_i|^2} \tag{2.2}$$

Being the goal to describe a conceptualized pose, one can consider a defined number of exemplars and compare the resulting FDs to the ones on a selected dataset, that are widely available nowadays, and consider the x closest matches.

FD is a quite good possibility as an effective descriptor that ought to store an extracted silhouette contours information. Specially considering an EDS approach, which would sample the extreme curvature points of the contours, which might potentially diminish the number of required sample points in order to effectively reproduce the extracted silhouette. However the fact that Fourier coefficients do not distinguish symmetry can become a problem when comparing extracted silhouettes.

2.1.2 - Shape Contexts

In [32] a basic SC is used to perform object recognition. In his work Belongie *et al*. successfully apply the algorithm and compare different shapes by matching each point to each other. In this case SC is calculated by given a set of discrete silhouette contour points n , with coordinates $p_i \in \mathbb{R}^2, i \in [1.n]$ obtained via an edge detector that represent a shape, the Shape Contexts algorithm consists on building an histogram of vectors that represent the distance relatively to a reference point of each one of the remaining $n-1$ points. Such histogram can be described by

(2.3). The log-polar binning purpose is making the descriptor more sensitive to neighbouring points rather than the farther away ones.

$$h_i^k = \#\{\, q \neq p_i : (q - p_i) \in bin(k) \,\} \tag{2.3}$$

SC ought to be used as a comparison tool between different video frames, as a mean to understand the action of the subject. The cost of comparing two points in different silhouette contours is described by (2.4) which signifies the quadratic distance between the SCs. p_i and q_j are points in different shapes being h_i and h_j their correspondent SC.

$$C_{ij} \equiv C(p_i, q_j) = \frac{1}{2}\sum_{k=1}^{K} \frac{|h_i(k) - h_j(k)|^2}{h_i(k) + h_j(k)} \tag{2.4}$$

Nevertheless a good result was achieved in the latter research, Belongie *et al*. continued the line of investigation and [16] a slight improvement in terms of efficiency was achieved by comparing using $\mathcal{L}^2 - norm$. Furthermore, two algorithms capable of retrieving the most probable candidates for comparison from a wider set of collected shapes are contemplated. This process is called fast pruning. The first method, Representative Shape Contexts consists on considering only a few SC from the selected shapes for matching by selecting the most suitable candidates taking in consideration their $\mathcal{L}^2$ distances as is illustrated in Figure 2.6.

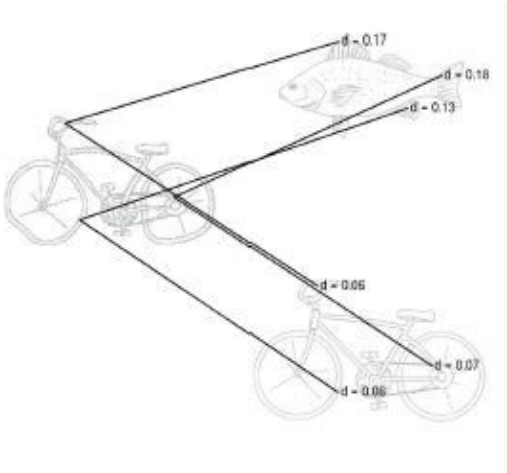


Figure 2.6: Representative Shape Context matching by measuring $\mathcal{L}^2$ distances (Source: adapted from [16])

On the other hand, when it comes to human silhouettes, one could foresee that in cases involving similar shapes although from different individuals, this method should not present itself very effective, especially if most considered shapes are human.

The second method, Shapemes [16] take advantage of vector quantization in order to lower the computational complexity of the calculations. It basically gathers SC histogram vectors into clusters which are called shapeme, diminishing the complexity of the overall SC.

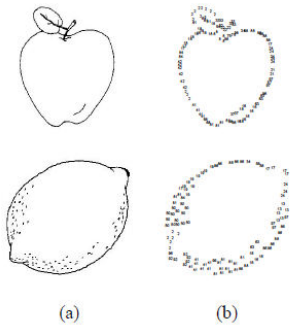


Figure 2.7: SC Shapeme algorithm (a) Original shapes, (b) Shapes after applying the Shapeme algorithm (Source: adapted from [16])

In Figure 2.7 it is visible the cluster division of the shape. The contour is basically sampled so that not such a big number of SC ought to be calculated.

SC descriptor might present itself as a valid alternative as a descriptor for a given shape. Particularly the Shapeme variation of the algorithm suggests an improved efficiency. It has the advantage of being sensitive to local limb orientation since the previously referred log-polar binning get finer in local regions.

2.1.3 - Hu moments

Firstly proposed by Hu in 1962 in his work [33], he divided the shape in two-dimensional moment invariants. However this first work using HU was not about human contour analysis, it was conceived for planar geometric figures analysis. The results showed successful recognition of successful patterns and alphabetical characters. A modified version of algorithm was proposed by Chen *et al.* in [34], in which he proposed an approach that on the contrary of the previous ones was merely based on the shape boundaries instead of all the shape itself (*i.e.* including the space inside it) ensuing, as a consequence, improvements in efficiency with comparable accuracy in recognizing sets of shapes, as reported in [35] where a comparative study over the two versions of the algorithm was conducted.

However, there are reports [36] stating that HU descriptor does not present satisfactory results comparing to other descriptors since it has two major downfalls, it is prone to lose local features (*e.g.* limbs) and likewise FD, it does not distinguish neither symmetric or rotated silhouettes.

2.1.4 - Distance Transform

As a discriminative method, DT describes a given shape by registering the distances between the contour points and other points on the silhouette input image. Such distance is denoted in (2.5), as

it was used in [37], where Elgammal *et al.* attempts to recover the body configuration and inferring a 3D pose directly from a human silhouette.

$$y(x) = \begin{cases} d_c(x) & if\ x\ is\ inside\ the\ contour\ c \\ 0 & if\ x\ is\ on\ the\ contour\ c \\ -d_c(x) & if\ x\ is\ outside\ the\ contour\ c \end{cases} \quad (2.5)$$

The distance $d_c(x)$ is the length to the closest point on the contour c. If such distance is negative the point is outside the contour, if it is equals to zero, it is on the contour, and if it is positive, the point ought to be inside the contour. The algorithm showed itself useful specially in contexts that aim to estimate body poses, as in the already referred work by Elgammal *et al.*

Despite it has produced satisfactory results in silhouette extraction [36], the algorithm lacks of efficiency compared to OM.

2.1.5 - Contour Signature

This type of descriptors usually store a certain geometric quantity of the shape contour sampled points, thus also being a discriminative method. Such geometric quantities include the complex coordinates along the contour relatively to the shape centroid, the Euclidian distances of the latter or the tangent angles of silhouette contour extracted points. CS algorithms are not rotational invariant since they require a starting point on the shape, usually the pixel situated at the most elevated point of the silhouette. Nonetheless, it can be achieved by computing several CS descriptors for the same shape with different starting points, as performed in [38] but obviously with an increased cost when it comes to computational effort. On the other hand, from the point of view of human pose analysis, this issue is to be taken into consideration because of the non-rigid nature of the human body. In [39], an early research work by Arkin *et al.* using CS, the algorithm is successfully applied in polygon shape analysis by using the angle of the counter-clockwise tangent as a function of the arc-length measured from a reference point on the boundary. More recently in [40], Rowe incorporates CS in an human pose tracking context.

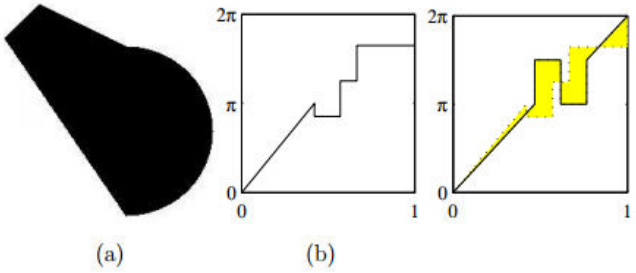


Figure 2.8: Shape contour tangent measuring as a CS (a) The retrieved shape, (b) the measured tangents, obtained counter-clockwise (Source: adapted from [40])

One of the heuristic measures used was the turning angle metric which measures the integral of the difference between two normalized functions that are derived from the tangents of each point of the silhouette contour.

In Figure 2.8 it is illustrated a tangent tracking of a shape contour which serves as input for the referred turning angle heuristic method that ought to describe a given silhouette.

CS could be a good alternative for silhouette representation. Despite its performance may be worse than other addressed algorithms *viz*. OM and DT [36], it is one of the most interesting options available.

2.1.6 - Poisson Features

As the name of the descriptor suggests, it takes advantage of the Poisson equation in the form presented in (2.6) to estimate the mean time required for a selected set of particles at a given point to, while performing a random walk, hit the contour. In (2.6), ΔU stands for the Laplacian of U whereas (x, y) are coordinates of the considered point and it ought to be noted that for the shape contour, $U(x, y) = 0$. Features such as polarity and orientation can be extracted and moments based on them are computed thus forming a shape descriptor.

$$\Delta U(x, y) = -1 \tag{2.6}$$

In [41], Gorelick *et al*. apply Poisson Features to describe silhouettes and further shape classification. The algorithm is then tested by using the extracted properties obtained after using the Poisson equation to successfully classify silhouettes retrieved from a public dataset.

Solutions to the Poisson equation provide rich descriptive information noting that a threshold can be used to decompose an object into parts. It is also stated [41] DT or other scalar fields do not present results as rich as PS.

2.1.7 - Discrete Cosine Transform

DCT can have variations within the equation describing it. Reid *et al*. in [42] postulated that given an image with size PxQ defined by $I(x, y)$, the DCT coefficients $M_{m,n}$ can be described as in (2.7), being $f_m(x)$ defined in (2.8).

$$M_{m,n} = \sum_x \sum_y f_m(x) I(x, y) f_n(y) \tag{2.7}$$

$$f_m(x) = \sqrt{\frac{1+\min(m,1)}{P}} \cos\left\{\frac{m\pi}{P} . \left(x + \frac{1}{2}\right)\right\} \tag{2.8}$$

According to the evaluation study in [42], the results showed that the more coefficients were calculated, the more precise would be the representation of the silhouette. In the referred study, DCT was compared to two other shape descriptors, including SC that is introduced in Section 2.1.2. Considering induced Gaussian Noise in the silhouettes used for the experiments *i.e*. the shape boundaries were typically affected, the results showed that although there is not a substantial difference, DCT outperformed SC. The latter is explained by the author by pointing out that DCT low order coefficients encode lower frequencies hence suppressing the noise.

According the results obtained by Reid *et al.* [42], DCT slightly outperformed SC when evaluated as human shape descriptors. The authors argue that DCT has an increased capacity of suppressing noise since the lower order DCT coefficients encode the lower frequencies.

2.3 - Skeletonization

Skeletonization presents itself as a further processing technique that acts upon a human silhouette contour as input. Its objective is primarily the detection of body parts thus providing a crucial role on further human action recognition. One of the first occasions this concept was applied successfully was in [43]. The authors proposed a start skeleton to represent the human shape, hence using it for further motion analysis. The principle was attaining the human center of mass and symbolise it as a star. With the knowledge of the contour points, the correspondent distance to the center of mass would be calculated accordingly thus extreme contour points would be detected as peaks in the function. The purpose of this work was to detect simple actions such as walking or running. Nonetheless the algorithm relieved itself as somewhat faulty when it came to correct differentiation of some parts of the body such as the head and the limbs in some cases. Following the same line of investigation, it was proposed by Aggarwal in an algorithm designed to detect people climbing fences [44] a two Start Skeleton (SS) model being the second star positioned at the higher extracted point of the silhouette. Afterwards, the calculated distances from each star were averaged in order to achieve greater precision. Later Chen *et al.* used another star skeleton model in [45], but additionally it was introduced an innovative way that permitted better results on obtaining the skeleton model and subsequently in the action recognition step of the respective work. A distance function between two SS was introduced, noting that each extracted SS was converted to a vector. The referred function was simply the sum of the Euclidean distances between five matched pairs which is used in an action recognition system based on the Hidden Markov Model.

Leveraging from its grounding efforts, Aggarwal achieved a breakthrough over these initial research works in [46]. The simple SS model had evolved to a Variable Star Skeleton representation, which, as it was demonstrated in the paper, showed better results on detecting points of extreme curvature, which was the major downside of the previous methods. The algorithm consists in designing an axis in the middle of the silhouette and consider several stars along it. For each star a set of extreme points is calculated and finally a selection of the extreme points is performed taking into consideration factors like robustness, the smoothing parameter, and most importantly, visibility and proximity from the star to the contour point.

In Figure 2.9 there are 4 stars represented by blue asterisks of which correspondent detection of the left hand of the subject is represented by a numbered red square. Clearly the fourth star showed the better detection, as the star number 2 showed a slight error. Stars 4 and 2 are in a better position relatively to the left hand when comparing to stars 1 and 3 simply because they have a clear line of sight to it. The second star, however, is far more distant from the hand than

the fourth, ergo showing that from the factors mentioned above, the proximity and visibility are clearly the most important.

This realization was what motivated Aggarwal in [44] to add an extra star to the model, so that a wider number of points of the silhouette contour would be visible to them thus providing a better approximation. However, an average of the distances from the two stars to the contour points is considered. Obviously there would be quite several points which would be visible to only one star of the two in which case the bad approximation would lead to an average point worse than the best approximation obtained. In order to explore this approach to its full, in [46] Aggarwal aims to use an appropriate number of stars and their corresponding positions so that as many points from the silhouette contour are visible to at least one star.

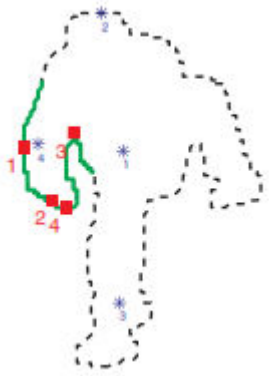


Figure 2.9: Hand detection from human silhouette using VSS algorithm (Source: adapted from [46])

The axis line can be computed using several methods [47]–[49] within which a threshold t is used to define how long each branch of the axis goes.

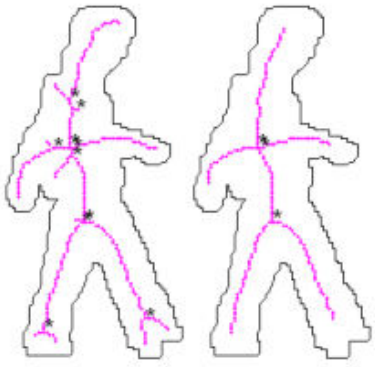


Figure 2.10: Medial axis computed using the augmented Fast Marching Method. On the left it was used the threshold $\boldsymbol{t = 10}$ and on the right $\boldsymbol{t = 30}$ is considered (Source: adapted from [46])

In Figure 2.10 it is represented a medial axis with different thresholds. The junction points are represented with blue asterisks since those ought to be the location of the stars (whenever they are closer than a defined threshold, both stars are merged and a mean point between them is considered as one). Assuming N stars are generated, the distance between $star_j, j \in [1, N]$ and each

point of the silhouette contour $P_i, i \in [1, NC]$ would be given by $dist_j(i)$ after which the prospective extreme points would be identified whenever a local peak was detected. Subsequently both the robustness and visibility are calculated. The first is measured as far away from the torso a point is, *i.e.* for a point k, an adjacent point k' on the opposite side of the torso, as it described in (2.9).

$$R = \frac{dist_j(k) - dist_j(k')}{|k - k'|} \tag{2.9}$$

The visibility V is hereby considered as the proportion of the line segment connecting the point to the star that generated it that is situated inside the silhouette. Once the calculations are done, the points with $R > MaxR$ and $V > MaxV$ are selected and the ones $R < MinR$ and $V < MinR$ are discarded, being $MaxR, MaxV, MinR$ and $MinV$ predefined thresholds. Finally only the top five extremity points are considered. If there are not enough selected points, then the highest classified points below the thresholds are selected in order to complete the desired number of extremity points.

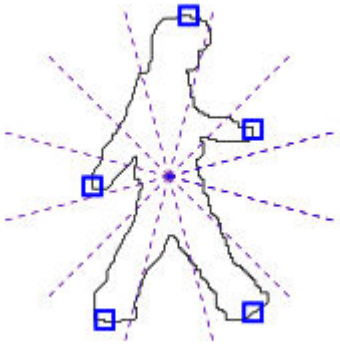


Figure 2.11: Histogram showing 12 feature vectors (Source: adapted from [46])

At this point the obtainment of the five major human body extremities is fulfilled and the silhouette ought to be sectored into N sectors, being the extremity points coordinates that are calculated accordingly to the center of mass of the subject. In Figure 2.11 is presented a human silhouette with 5 extremity points discriminated and it was divided in 12 parts. Taking this knowledge into consideration, one can consider the further action recognition steps. Knowing the evolution of the position of each part of the body over time, movements such as running, walking, squatting, can be categorized.

2.4 - Summary

Being the foreground segmentation a step prior to the application of the algorithm, Section 2.1 cover two methods that can be used to apply it. Background subtraction presupposes an ongoing scene with a still background whereas Saliency Maps are usually applied to still images. Subsequently, Section 2.2 cover some contour descriptors that can be used to represent the extracted silhouette. Those image descriptors can be used to effectively compare silhouette

contours, being in the video sequence in which they are extracted, for action recognition analysis, or for object recognition purposes *i.e.* detecting a given object (*e.g.* a human) in a general scenario.

Skeletonization is addressed in Section 2.3 as a quite valid option of pursuing the computing of the five feature points which are the output of the proposed algorithm in Section 3.2.

Chapter 3

Silhouette Extremities Identification

In this chapter, it is introduced the proposed method used to automatically extract and identify in video sequences the location of silhouette feature points from the head, left and right feet and hands (whenever it is referred the left and right foot or hand, it is considered to be the foot or hand on the left or right side of the image). In Figure 3.1 it is possible to see an example of a successful application of the proposed algorithm. As it is plain to see, in this case the error lines are quite small since this is a good example of an effective application of the algorithm. In this frame, not only the five feature points were detected, but also successfully identified as belonging to the head and the left and right hands and feet.

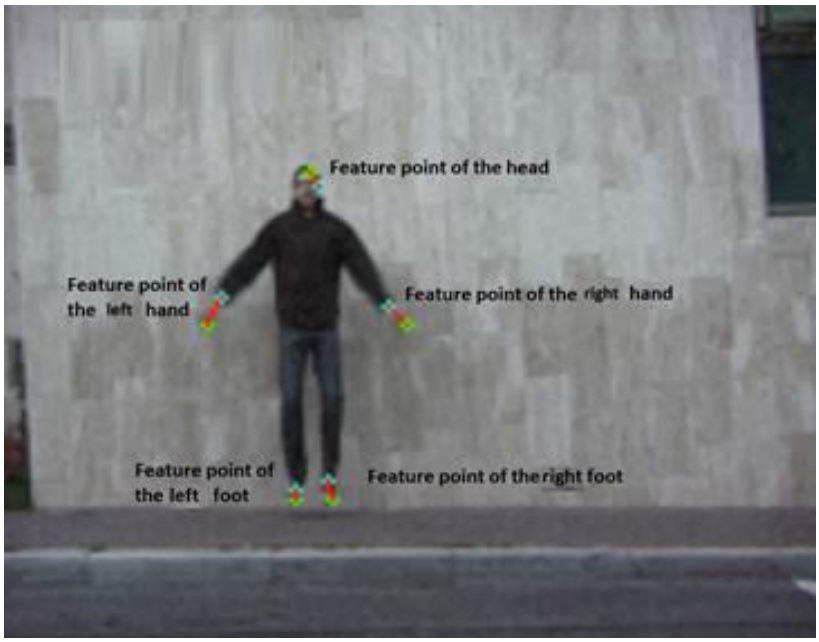


Figure 3.1: Frame illustrating a successful detection and identification of the feature points. The reference and detected feature points are represented in green and blue respectively. The red lines represent the distance error between them.

As the work developed by J.K. Aggarwal [44] served as an important base, in Section 3.1 an overview of the latter is provided in order to facilitate a further understanding of the contribution of this research work. When it comes to human silhouette representation, even though Aggarwal effectively detects body extremities feature points, the task of identifying from which major body extremities those feature points belong to, like hands, feet and head is not addressed since, according to the author, that information is not required for the goal of his work which is detecting humans climbing fences. Yet, that information may be relevant in order to broaden Aggarwal's work, and instead of merely detecting Humans climbing fences, it may be used to automatically classify simple types of movements. Thus, the major contribution of this research work is proposing a way of performing a correspondence between the detected silhouette's extremities feature points and the head, hands and feet of the Human body. Additionally, an alternative measure providing relative anatomic information of the silhouette is proposed namely four angles, referring to the hands and feet of the human body.

In Section 3.2, the proposed method for identification of the five human body feature points as well as the delimitation of the angles is thoroughly described.

As part of an exploration of alternative methods to describe a human silhouette contours, a modified version of the proposed approach is presented in Section 3.3. It combines the feature points obtained using the Freeman Chain Code based algorithm covered by Yao *et al.* [50] with the five feature points obtained from the proposed method addressed in Section 3.2 as a post-processing step in an effort to strengthen the identification results of the latter.

3.1 - Base method for feature point detection

The method described in this section was proposed by J.K. Aggarwal [44] as part of an overall system which included a classification phase mainly performed using a Hidden Markov Model which took as an input sequences of the detection of feature point's algorithm. However, for the purposes of this work, only the feature points' detection part of the algorithm is exposed. One of the purposes of this section is also provide a base of comparison for the proposed algorithm of Section 3.2, for a clearer understanding of its contribution.

The first task to achieve is to perform the foreground segmentation. In order to obtain the subject's silhouette to be represented, a background subtraction algorithm is applied. For the sake of simplicity, solely the biggest blob is considered, thus the assumption that only one person is considered for analysis in each scene is considered.

Resulting from the background subtraction, a set of contour points is obtained by attaining the frontier points between the background and foreground of the foreground mask. They are used as input to a following more refined silhouette representation. Firstly, two reference points, which the author names stars (the silhouette representation model is often referred to as star-skeleton model), are considered: the centroid of the contours and the highest point of the contours.

Afterwards, two vectors of Euclidean distances between each star and all the contour points are computed. A plot of these vectors is presented on Figure 3.2(b). The feature points to be considered are the contour points which present themselves as local maximums approximately in the same range of indexes in both distance vectors.

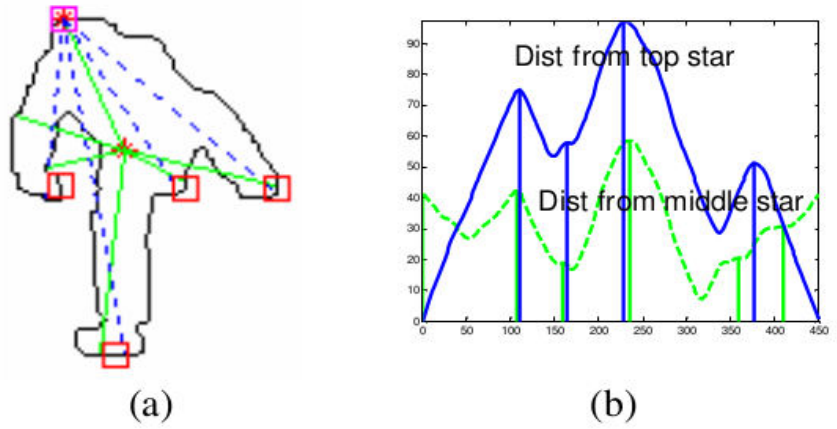


Figure 3.2: Two star-skeleton model (Source: adapted from [44]) (a) Extracted silhouette, two stars and feature points; (b) Plots of distances from each star and all contour points

In Figure 3.2(a), both the highest contour and centroid stars are represented along with straight lines connecting them with the detected feature points which are signalized with red squares. The author then uses simply criteria involving the number of feature points above of bellow the point of the fence and the location of the centroid to decide if the subject actually is about to climb over it. Even though it was not relevant for the application addressed by the author, one could think of other scenarios where the constant knowledge of the location of the feet, hands and head of a subject and other anatomic information relating their locations could be interesting. For instance, classifying courses of action automatically, like walking, running, jumping and from that point, one could build even more complex models, hence identifying sequence of more complex actions. This could be of great interest for large surveillance systems, potentially decreasing the number of people monitoring them.

In the next section, an approach for identifying the previously referred five silhouette feature points is presented, as well as the proposal of four angles involving them which offer a size invariant measurement that can present itself useful in describing the subject's movement.

3.2 - Proposed approach for feature point identification

It is important to stress that the main contributions of the algorithm to be presented are describing the human silhouette by identifying the head, hands and feet feature points as well as providing a measure relating them that contains information regarding the kind of movement being conducted by the subject, namely the angles described in Figure 3.5.

In Figure 3.3, a diagram describing the algorithm is presented. It was designed taking into consideration the environment in which it was tested (a sequence of videos from a dataset containing different types of movement) that is addressed in Chapter 4. Nevertheless, the foreground segmentation, even though it is the base of this process, was considered a granted part of it, given the fact that it has been broadly explored in the literature as it was attended in Section 2.1.

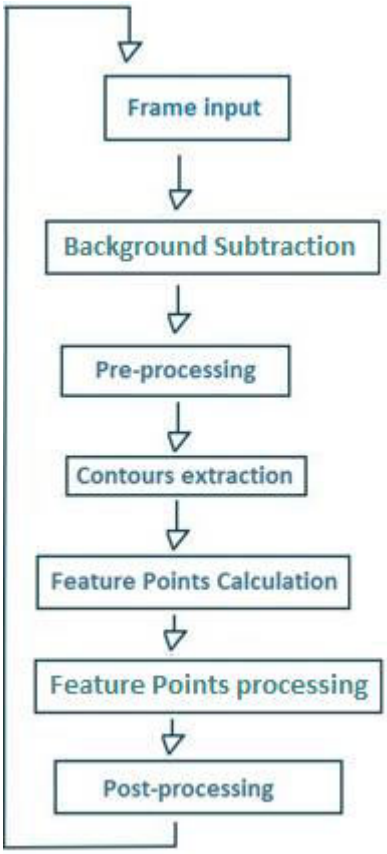


Figure 3.3: General diagram of proposed algorithm

First of all, after a video from the testing dataset is introduced as input, each frame is processed subsequently, starting with the computing of the foreground mask. The used background subtraction method was a Mixture Of Gaussians and the parameters included the frame history, the number of Gaussian mixtures, the background ratio and the noise strength. Afterwards, the pre-processing includes a simple binary threshold is performed in order to assure that all the pixels acquire values of 0 and 1 exclusively. This step is necessary because there were cases where the output of the background subtraction did not present a flawless foreground mask, it contained irregularities that ought to be eliminated. Also, two consecutive morphological operations, dilation and erosion were performed for the purpose of noise reduction of the collected masks.

As for the contours extraction, the frontier points between the foreground and background are computed and stored as the silhouette contour points. Naturally, independently of the fact that in this case there only was a single person in each video sequence, the problem of small noisy

foreground blobs was resolved by simply consider the largest set of contours. In a scenario containing many people, one could suggest the usage of a Human detection algorithm [51]–[53] *a priori* the application of the background subtraction where the location of the detected blobs containing humans could be compared in terms of size and location in the frame in order to determine if a given blob resulting from the background subtraction actually corresponded to a Human.

3.2.1 – Feature Points Matching

At this point of the algorithm, the vital information to compute the position of the desired five feature points is available, the contour points of the subject. Therefore, the "Feature Points Calculation" stage is conducted by firstly computing two vectors, similarly as explained in Section 3.1: one containing the Euclidean distances between the highest point of contours and all the contour points and the second the Euclidean distance between the centroid and all the points of the contour. However, instead of considering the local maximums that are located in roughly the same neighborhood in both vectors, it was chosen to sum the vectors and consider all the local maximums as candidates for the final five feature points. The plot of the summed distance vectors corresponding to the frame illustrated in Figure 3.1 is presented in Figure 3.4(a). Throughout this thesis, the y-values represented in this plot is referred to as combined distances, since they refer to the summed distances from the plots on Figure 3.2(b) referred in Section 3.1. In the y-axis, the summed distances are represented in pixels and local maximums were represented with red crosses. Though the local minimums are represented in violet stars on the plots, those were represented for a more detailed analysis and were not used in any part of the proposed algorithm. In this case, exactly six local maximums were detected, so these are the candidate feature points. The x-axis denotes the contour point that each combined distance corresponds to.

The feature point of the head is considered to be the nearest candidate point to the highest point of contours. As it is discussed in specific examples in Chapter 4, a few drawbacks were detected with this approach, however for most common positions, it is a valid assumption. Before continuing with the criteria for choosing the feature points of the hands and feet, let's establish that whenever the expressions right or left side of the centroid are used, considering the Cartesian axis centered in the Human centroid visible on Figure 3.4(b), the plan of every point with a higher abscissa than the centroid the right side it and the plan containing all the points with a lower abscissa than the centroid's, the left side of it. As for the feet, the sub-area where they are searched is below the centroid, as it shows Figure 3.4(b). Therefore, the feature point with the highest combined distance from the left side of the centroid is considered to be the foot on the left side of the frame and the feature point with the highest combined distance from the right side of the centroid is assigned as the foot on the right side of the image. This criteria is true for the

majority of cases since the feet are the feature points located further way from both the centroid and the highest point of contours.

The cases where no candidate feature points are detected on one of the sides of the centroid are covered by assigning to the missing foot the coordinates of the other. Consequently, it is assumed that given these criteria, at least one foot feature point is detected. This last condition is applied in order to include the cases where both feet are joined together, hence they would not clearly be qualitatively differentiated by the extracted silhouette nor would two local maximums with similar amplitudes be detected in the region considered for the feet. In Figure 3.4(b) it is observable that it is not the case, since the feet are well set apart from each other which naturally leads to two well defined local maximums in Figure 3.4(a) that actually correspond to the two highest maximums of the plot. However, in several poses it is reasonable to consider the same feature point for both feet, when they are not evidently identified by two local maximums. In Chapter 4, this situation is addressed and evaluated alongside with the matching results.

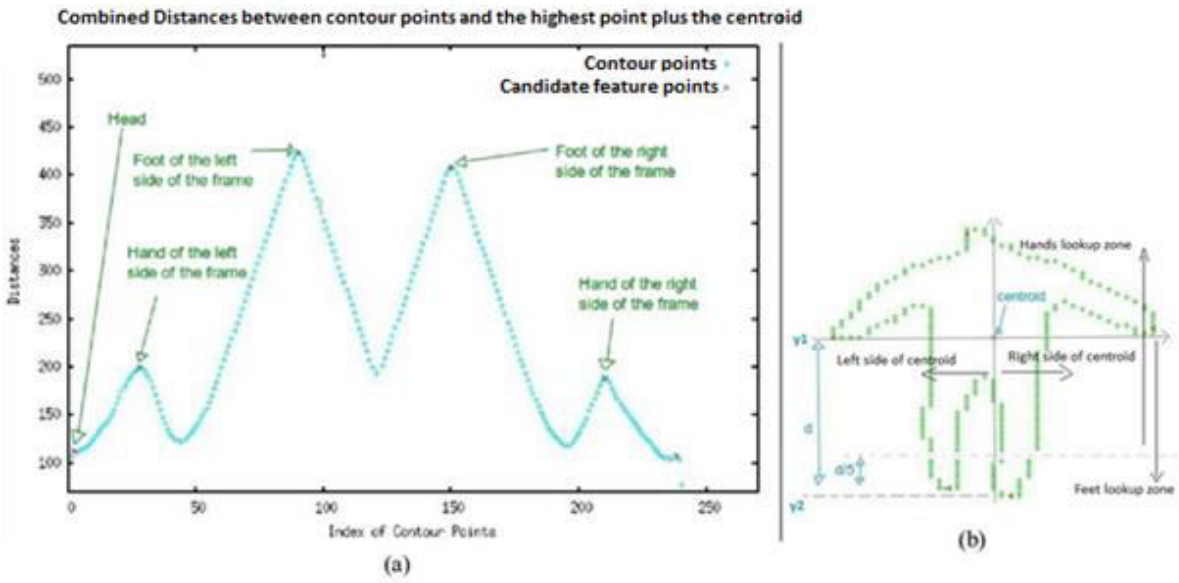


Figure 3.4: Criteria for feature point selection (a) Combined distances plot between highest point and centroid to all contour points; (b) Region delimitation on the contours

With respect to the identification of the feature points of the hands, firstly consider the zone where they are considered to be, the plane above $y2 + d/5$ represented in Figure 3.4(b): let $y1$ be the y-coordinate of the centroid and let $y2$ be the y-coordinate of the foot whose feature point has a lower y-coordinate. Finally let $d = |y1 - y2|$. The hand on the left side of the image is considered to be the highest local maximum of the combined distances vector with a y-coordinate higher than $y2 + d/5$. The same criteria applies for the hand on the right side of the image, instead in this case the criteria is narrowed to the right side of the centroid. For the majority of poses, the two feet would correspond to the two highest local maximums of the combined distances plot and the feature points of the hands were positioned nearby the third and fourth highest peaks. Thus, a

reasonable criteria would be consider the hands to be the third and fourth highest local maximums of the combined distances plot. However, taking in consideration the irregularity of the contours provoked by noise resulting from the background subtraction, more than one local maximum could be detected in the neighborhood of both feet. Hence, by assuming that the hands don't undertake positions below $y2 + d/5$, condition that for most common poses is confirmed, these noise originating irregularities are avoided.

3.2.2 - Noise reduction

The background subtraction may sporadically cause noise, which may be clearly identified by verifying a spontaneous increase in the distance between the reference and the detected feature point. In order to eliminate these errors, the five feature points of the previous frame were taken into consideration.

Let the height of the subject be defined by h in pixels. Let i be the index of the five detected feature points, let n be the index of the frame of a given sequence of frames and let N be the total number of frames in the sequence. Also, let $f[i][n]$ be the coordinates of the feature point i of the frame n. Consider the Euclidean distance (in pixels) between the same feature point in consecutive frames defined by $e[i][n]$. So, $i = \{1..5\}$, $n = \{1..N\}$ and $e[i][n] = ||f[i][n-1] - f[i][n]||$. For $n >= 2$, if $e[i][n] > h/2$, then $f[i][n] = f[i][n-1]$. The threshold $h/2$ was chosen taking in consideration individual frame analysis and it was concluded that the motion carried on the videos did not permit that in consecutive frames, a feature point would not move more than the referred threshold. This noise-reduction step permitted the correction of irregularities of some detected feature points which consequently led to better overall results.

This condition mainly entails two assumptions. Firstly, that these kinds of errors do not occur in consecutive frames. And secondly an error of this magnitude does not occur on the first frame, otherwise it could potentially create a chain error deterioration. This condition was added to the algorithm because of the analysis of the results and it did correct major detection errors in several frames. A trade-off should be reflected over these factors taking in consideration the environment in which it is applied. If this noise error occurs in consecutive frames, then instead of merely considering the previous set of feature points, it should be considered the set of feature points of the previous k frames.

3.2.3 - Angular Descriptor

The stage "Feature point processing" refers to the computing of the data relative to each frame that ought to be used for statistical processing by the end of the video and also the calculation of the 4 proposed angles. In Figure 3.5 it is possible to see the 4 proposed angles that are calculated using basic trigonometric analysis, as presented in formulas (3.1)-(3.4).

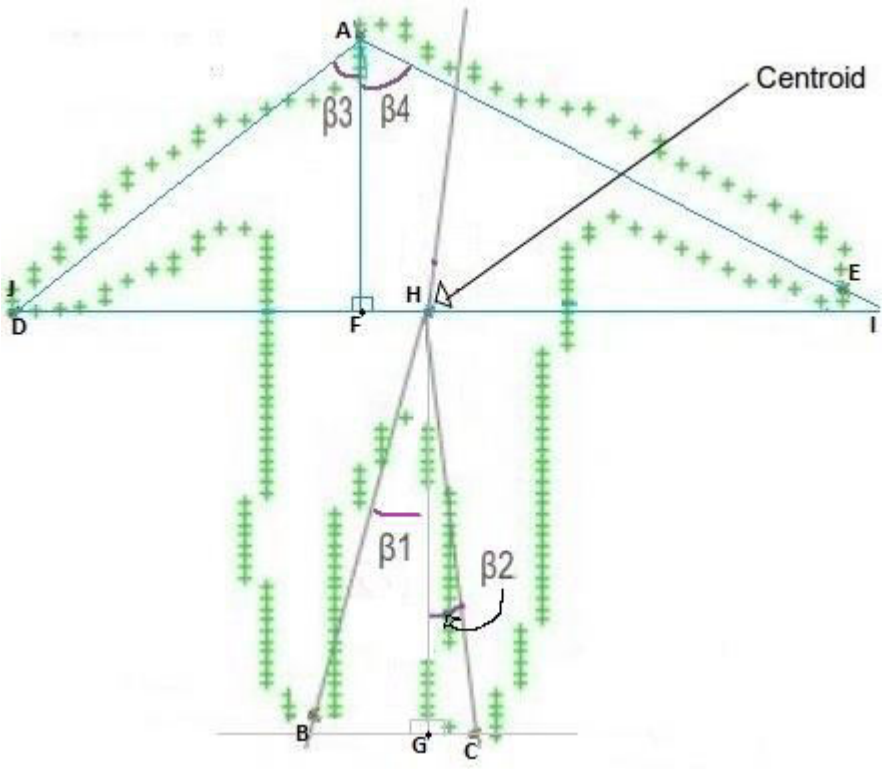


Figure 3.5: Proposed angles denoted on extracted contours from the Human silhouette in Figure 3.1. A..E: Feature points of the head, left foot, right foot, left hand and right hand, respectively; H: centroid of contours; F: intersection of straight lines AF with DI; G: intersection of straight lines GH and BC; I: intersection of straight lines IF and AE; J: intersection of straight lines AD and DF; $\boldsymbol{\beta 1 = B\hat{H}G}$; $\boldsymbol{\beta 2 = G\hat{H}C}$; $\boldsymbol{\beta 3 = D\hat{A}F}$; $\boldsymbol{\beta 4 = F\hat{A}E}$

$$\boldsymbol{\beta 1 = \cos^{-1}(\overline{GH}/\overline{BH})} \quad (3.1)$$

$$\boldsymbol{\beta 2 = \cos^{-1}(\overline{GH}/\overline{CH})} \quad (3.2)$$

$$\boldsymbol{\beta 3 = \cos^{-1}(\overline{AF}/\overline{AJ})} \quad (3.3)$$

$$\boldsymbol{\beta 4 = \cos^{-1}(\overline{AF}/\overline{AI})} \quad (3.4)$$

It should be noted that these angles were considered as a descriptor that takes advantage of the identification of the 5 extremity points. The angles $\beta 1$ and $\beta 2$ aim to relate the position of the hands relatively to the head. On the other hand, $\beta 3$ and $\beta 4$ relate the position of the centroid with both feet. They can also be analyzed in order to delineate the position of each foot relatively to the other.

These angles may be an interesting input feature to a classification system, since they provide scale-invariant data which might present certain patterns for a given type of movement, which is demonstrated in Section 4.6.

It is worth to emphasize that other measures relating the 5 feature points can be considered. The angles $\beta 1, \beta 2\, \beta 3$ and $\beta 4$ are just an example that illustrates the potential of the identification of the proposed 5 feature points towards the proposal of other features relating them.

After the processing of the feature points, the post-processing would take place now. This phase is addressed in Section 3.3. It is considered as part of an extended proposal of the algorithm noting that it is evaluated with and without it in order to assert its impact.

3.3 - Post-processing

As an alternative way of providing feature points, it was considered a different approach which presented a different set of results than the proposed five feature points output. This alternative approach however, was used as a way to improve the detection of the five feature points obtained using the proposed algorithm in Section 3.2. It was implemented based on the proposal of Yao *et al*. [50], with a slight amendment that is attended later on, to the algorithm presented by the author. The feature points outputted by this algorithm have the potential to provide a more comprehensive anatomical analysis. The rational of the proposal of its integration with the proposed algorithm is that by knowing the anatomical area around each of the 5 detected feature points, it may be possible to improve its location by reassigning it to a position more likely to be accurate.

In this section, the proposed implementation of the latter algorithm is presented as well as how the resulting feature points were combined with the five feature point output. It is also important to add that the proposed algorithm was tested with this post-processing step and without it, in order to evaluate if it improved the results and in which conditions. Even though no substantial improvements were verified, the post-processing outputs contour points that can allow an understanding of each of the 5 feature points surrounding anatomy. The criteria used to combine the post-processing with the proposed algorithm can indeed be improved in order to take advantage of this fact.

Firstly, the Freeman chain code [54] is computed for every point of the contours, attributing to each point a number between 0 and 7, which would be decided taking into consideration the location of the next pixel of the contours, thus defining a direction as illustrated in Figure 3.7.

Considering a mathematical notation, let N be the number of points of a given set of contours C, and let i be the index of a given point of the set of contours. $i = \{\, 0..N \,\}$. Let $C[i]$ represent the point of index i of the set of contours C and let $C[i].x$ and $C[i].y$ represent the x and y-coordinate of the given point. Consider at last that the Freeman chain code of a given point is assigned to $C[i].freeman$. The Freeman chain code of $C[i]$ was defined based the conditions represented in (3.5-(3.12).

$$\text{If } (C[i+1].y == C[i].y) \;\&\&\; (C[i+1].x >= C[i]) \text{ Then } C[i].freeman = 0 \tag{3.5}$$

$$\text{If } (C[i+1].y >= C[i].y) \;\&\&\; (C[i+1].x >= C[i]) \text{ Then } C[i].freeman = 1 \tag{3.6}$$

$$\text{If } (C[i+1].y >= C[i].y) \;\&\&\; (C[i+1].x == C[i]) \text{ Then } C[i].freeman = 2 \tag{3.7}$$

$$\text{If } (C[i+1].y >= C[i].y) \;\&\&\; (C[i+1].x <= C[i]) \text{ Then } C[i].freeman = 3 \tag{3.8}$$

$$\text{If } (C[i+1].y == C[i].y) \,\&\&\, (C[i+1].x <= C[i]) \text{ Then } C[i].freeman = 4 \tag{3.9}$$

$$\text{If } (C[i+1].y <= C[i].y) \,\&\&\, (C[i+1].x <= C[i]) \text{ Then } C[i].freeman = 5 \tag{3.10}$$

$$\text{If } (C[i+1].y <= C[i].y) \,\&\&\, (C[i+1].x == C[i]) \text{ Then } C[i].freeman = 6 \tag{3.11}$$

$$\text{If } (C[i+1].y <= C[i].y) \,\&\&\, (C[i+1].x >= C[i]) \text{ Then } C[i].freeman = 7 \tag{3.12}$$

Afterwards, line segments are conceptualized given the directions associated to each point *i.e.* if $C[i].freeman == C[i+1].freeman$, then $C[i]$ and $C[i+1]$ would belong to the same segment which would have a number between 0 and 7 associated, extrapolating the rationale behind the attribution of that same number to an individual point, depending on the direction of it. While evaluating the sequence of points of a silhouette, whenever the direction of a point changed, a new segment was formed. Figure 3.6 illustrates consecutive line segments and the directions associated with each one of them.

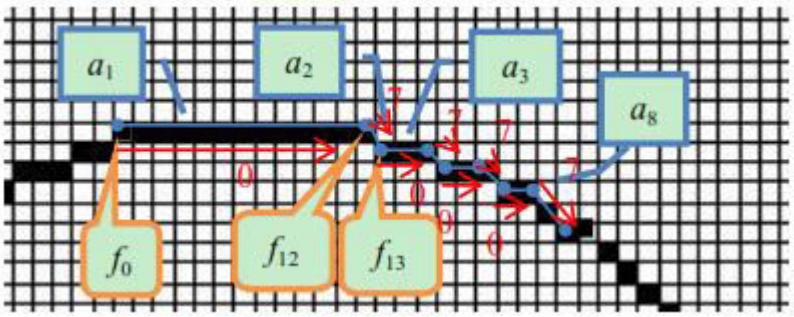


Figure 3.6: Example of line segments and correspondent freeman chain code associated (Source: adapted from [50])

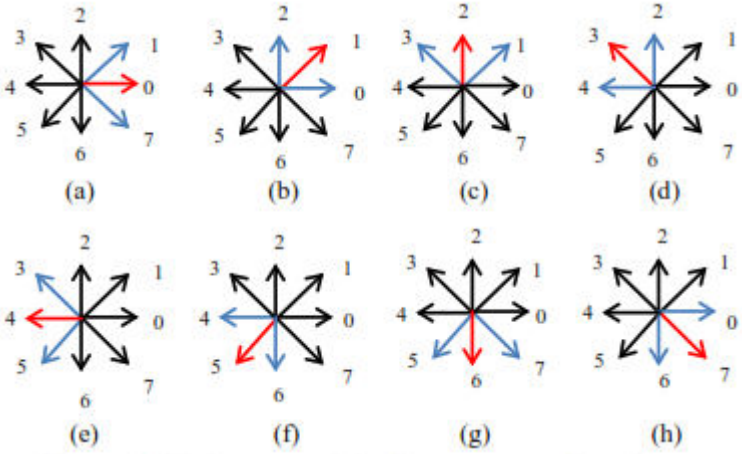


Figure 3.7: Possible directions to be considered for Freeman chain code (Source: adapted from [50])

$$\begin{aligned} &\text{(a) } d_2\text{-}d_1\text{=-1},d_2\text{-}d_3\text{=1} \quad \text{(b) } d_2\text{-}d_1\text{=1},d_2\text{-}d_3\text{=-1} \\ &\text{(c) } d_2\text{-}d_1\text{=-1},d_2\text{-}d_3\text{=-7} \quad \text{(d) } d_2\text{-}d_1\text{=-7},d_2\text{-}d_3\text{=-1} \\ &\text{(e) } d_2\text{-}d_1\text{=1},d_2\text{-}d_3\text{=7} \quad \text{(f) } d_2\text{-}d_1\text{=7},d_2\text{-}d_3\text{=1} \\ &\text{(g) } |d_2\text{-}d_1|\text{=2} \end{aligned} \tag{3.13}$$

$$|d2 - d1| = 6 \tag{3.14}$$

Being $d1$, $d2$ and $d3$ the freeman chain codes of consecutive line segments, the equations on (3.13) are postulated in order to calculate the output feature points. Note that these conditions are already illustrated in Figure 3.7, where $d2$ would correspond to the freeman chain code of the red segment. The matching between the formulas and the figures is the following:

- Formulas (3.13)(a-b) match Figures 3.7(b-g);
- Formulas (3.13)(c-d) correspond to Figure 3.7(a);
- Formulas (3.13)(e-f) match the Figure 3.7(h).

Additionally, the author considers (3.13)(g) as the case where two line segments are perpendicular. However, if all the perpendicular line segments are to be considered, another condition should apply. So, it is proposed in this implementation to also consider equation (3.14) on the computing of the feature points. This equation is valid in the cases where consecutive line segments acquire one of the combinations of the freeman chain code values represented in (3.15) and (3.16).

$$d1 = 1, d2 = 7 \; or \; d1 = 7, d2 = 1 \tag{3.15}$$

$$d1 = 0, d2 = 6 \; or \; d1 = 6, d2 = 0 \tag{3.16}$$

These values would mean the line segments were also perpendicular, case that was not covered by the author.

Finally, having computed the feature points, the latter are submitted to a pruning process in order to refine the selection *i.e.* the feature points whose Euclidean distance to its predecessor is lower than a threshold are excluded. The referred threshold was based on the pixel resolution of the videos. The lower the threshold, the higher is the anatomic information that the whole set of feature points can provide.

The five feature points outputted by the proposed algorithm in Section 3.2 are hereby substituted by the nearest feature point calculated from this algorithm. A couple of things should be taken into consideration for this criteria:

- The set of feature points originated from this Freeman chain code based algorithm tend to originate a higher number of points, hence it is likely that one of them would be in a narrow neighborhood of each of the original five feature points;
- It can though provide a simplistic way of describing the anatomic area nearby each of the original five feature points, ergo possibly validating or not its correct detection. However, this paradigm was not deeply explored because of work schedule reasons.

Evidently, alternative ways could be postulated in order to improve the detection of the five body feature points by combining their original locations with the data provided by the Freeman

chain code based algorithm. It is suggested as a further investigation path that could lead to improved results of the original proposed algorithm in this work.

3.4 - Summary

This chapter mainly served to explain the proposed algorithm that was successfully implemented. Firstly, a contextualization has been made by briefly addressing the detection algorithm that served as a base for the 5 feature point identification of the Human body. Afterwards, details about the formulation and development of the suggested procedure have been provided and additionally, a post-processing method has been proposed. Furthermore, an overall explanation of the purpose of not only the algorithm itself, but also why it is relevant to propose it has been provided.

Chapter 4

Experiments and results

The experiments and corresponding results in which the algorithm was testes are presented in this Chapter. The results are discussed and analyzed from different perspectives and the algorithm is evaluated accordingly.

The algorithm was tested with the classification dataset "Weizmann" [55]. It contains sets of short term videos with 9 different people in the same background scenario though with illumination and contrast variations as well as different clothing of the subjects. 10 different actions are conducted which are illustrated in Table 4.1.

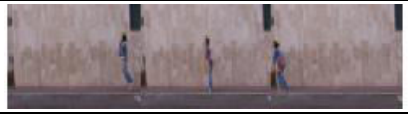
Action	Description	Actions	Description
Bend		Side	
Jack		Skip	
Jump		Walk	
Pjump		Wave1	
Run		Wave2	

Table 4.1: Illustration of the 10 actions of the "Weizmann" dataset [55]

The experiments and the conditions in which they were conducted are detailed in Section 4.1. The matching of the detected feature points and the reference ones is a major aspect of analysis in Section 4.2. The results considering the detected feature point nearest to each reference feature

point are evaluated in Section 4.3. Moreover, the proposed post-processing method is discussed alongside the corresponding results in Section 4.4. In Section 4.5 it is addressed the impact of background subtraction on the attainment of an accurate foreground segmentation and to what extend it influences a good performance of the algorithm. Additionally, in Section 4.6 the proposed 4 angles are evaluated both from the obtained error perspective and also as a mean to provide data that can be associated with specific simple actions like the ones performed on the testing videos.

For the purposes of discussing the results in the proposed perspectives in this chapter, only the most relevant generated plots is presented. However, a more extended part of the graphical results are exposed in the Annexes of this document.

4.1 - Experiments Setup

Even though the dataset offered 9 different people performing the same action, only 3 videos per action (totalizing 30 videos) were manually processed in order to obtain the reference location of the five feature points per frame because it was considered that it was enough to perform the desired analysis. For the purpose of the description and discussion of the experiments, the 5 points obtained through this procedure are referred to as reference feature points.

4.1.1 - Annotated data

On the matter of the verification of the obtained results, since the goal of the proposed algorithm is to identify 5 extremities of the body, the coordinates of the desired 5 five feature points were identified for each frame of each examined video. For that purpose a script has been developed using Matlab that enables reading a video and store in files the coordinates of the 5 points manually identified with the mouse in each frame. In Figure

4.1 the process of the reference feature point marking is illustrated. As stated, it was conducted for 30 videos which in average have nearby 100 frames, totalizing approximately 3000 frames and 15000 points.

4.1.2 - Evaluation measures

Each of the reference feature points was compared with the output of the proposed method and error measures were conducted. Essentially, three errors were calculated:

- The point-to-point Euclidean distance between each detected and reference feature points;
- The point-to-point Euclidean distance between the reference and the nearest detected feature points;
- The difference between the proposed angles illustrated in Figure 3.5 obtained using the detected feature points and the same angles calculated with the reference feature points.

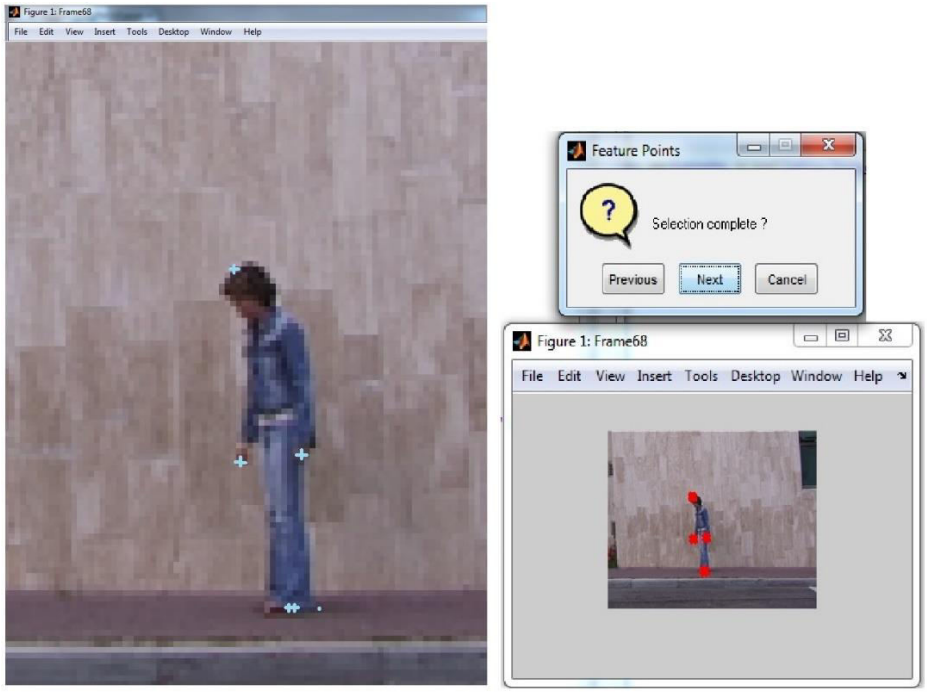


Figure 4.1: Identification of reference feature points using Matlab a) Manually identified points with mouse; b) Confirmation window that enables the remarking of points

The precision was also measured for each video. For that matter, it was established that every detected feature point whose Euclidean distance to its corresponding reference location was lower than a defined threshold, was considered a true positive, otherwise it would be a true negative. Table 4.2 contains the detected and reference coordinates of each feature point for the frame presented in Figure 3.1, as well as the Euclidean distance between them symbolizing the error, which being lower than the defined threshold, leads to an effective identification of the corresponding point.

Feature Point	Euclidean distance between reference and detected points (in pixels)	Threshold = 15% of the height of the subject (in pixels)	Successful identification
Head	3.61	10.95	YES
Left Foot	2.00		YES
Right Foot	4.12		YES
Left Hand	6.70		YES
Right Hand	5.66		YES

Table 4.2: Evaluation of the identification of the five feature points for Figure 3.1

The results presented in this document were generated for a threshold defined upon an approximation of the size of the head based on the subject's height. Hence, the subject's height in pixels was calculated by considering the difference between the y-coordinate of the reference head and the lowest positioned foot feature point in a frame of every video where the subject was

standing. Additionally, the Euclidean distance errors presented in the plots along this Chapter are all normalized relatively to the height of the subject.

Furthermore, for each of the experimental scenarios referred in Section 4.1.3, the outputted evaluation measures for each feature point for each video is the enumerated in Table 4.3. On the other hand, after the processing of all the videos, the same data (except the per frame analysis of each measure) was averaged per category of action and for the total number of videos in order to get a bigger picture of the overall performance and its limitations according the type of action carried out.

<table>
<tr><th>Angles</th><th>Euclidean distance error</th><th>Precision</th></tr>
<tr><td>Average angle errors</td><td rowspan="2">Average distance errors</td><td rowspan="2">Precision for overall video</td></tr>
<tr><td rowspan="2">Standard deviation of angle errors</td></tr>
<tr><td rowspan="2">Distance error per frame</td><td rowspan="3">True positives/negatives per frame</td></tr>
<tr><td rowspan="2">Angle error per frame</td></tr>
<tr><td rowspan="2">Standard deviation of distance errors</td></tr>
<tr><td>Reference angle per frame</td></tr>
</table>

Table 4.3: Evaluation measures for each feature point for each video

Moreover, during this Chapter, many plots are presented regarding several evaluation measures of the feature points. In these plots the following nomenclature is adopted to the feature points:

- Foot1 corresponds to the foot which is most on the left of the frame;
- Foot2 corresponds to the foot which is most on the right of the frame;
- Hand1 corresponds to the hand which is most on the left of the frame;
- Hand2 corresponds to the hand which is most on the right of the frame;

4.1.3 - Experimental scenarios

Four top level input parameters that constrained the output of the experiment were defined:

- The usage of the ground-truth masks as input video or application of background subtraction (MOG) to the original videos in order to obtain the foreground segmentation;
- The application of the post-processing or not;
- Utilization of the detection-only version of the algorithm;
- The threshold used for precision calculation.

Detection-only version of the algorithm refers to the adoption of an error calculation paradigm which considers for each of the five reference feature points, the nearest of the five detected in terms of Euclidean distance. For instance, consider the case where the Euclidean distance between the reference feature point of the right hand and the detected one is 20 pixels. However, the

lowest Euclidean distance of any of the detected points and its reference right hand point is the detected left hand point which presented a 15 pixels distance. The algorithm would then select the detected point for the left hand for error measuring relatively to the reference right hand feature point. This version of the algorithm measures how well the five feature points are detected, independently of which feature point they intent to identify.

Taking into consideration these parameters many experimental scenarios were tested. However, the following four were considered in order to perform the desired analysis to the algorithm's performance:

1) The usage of background subtraction on the algorithm without post-processing;
2) Considering the ground-truth videos as input on the algorithm without post-processing;
3) Considering the ground-truth videos as input on the algorithm with post-processing;
4) Considering the ground-truth videos as input on the detection-only version of the algorithm;

These scenarios are used to analyse the results in different perspectives, namely the impact of background subtraction, which is assessed considering the results obtained from scenarios 1) and 2). The effect of post-processing is addressed comparing scenarios 2) and 3) and the detection performance is evaluated independently of the matching results by comparing the output of scenarios 3) and 4).

4.2 - Point matching evaluation

In this section, the general performance and the limitations of the algorithm that apply to the different experimental scenarios are evaluated. First of all, assuming the ground-truth videos contain foreground masks completely free of noise, they are considered the reference input for the algorithm. Therefore, for the purpose of a general evaluation of the algorithm, the precision results studied in this Section are the ones obtained from the experimental scenario where ground-truth videos were used.

Consider the Figure 4.2 where the precision averages of the five feature points for all the videos are presented. It is clear the robust results that the head and both feet present. The average precision rates on all the videos of both feet and head feature points is higher than 90%. On the other hand, the precision rates of both hands are not so satisfactory, even though the average Euclidean distance error for both of them is around 30% of the height of the subject, as it is visible in Figure 4.3. However, the average higher error of hands' feature points is explained by the fact that for most of the actions on the dataset, the shape of the hands is occluded on the inside of the silhouette contours which did not permit a good performance. A more detailed analysis on this matter is exposed during this section.

The highest precision rates are verified in "wave2", "wave1" and "jack" actions as it is visible in Figure 4.4. The reason why this happens is because "jack" is an action where most of the time, the limbs of the subjects are well set apart from the torso and on favorable positions for a good

detection *i.e.* one hand and foot in each side of the centroid. Generally in every action, feet and head have quite good scores as it can be concluded by Figure 4.2 which shows the average precision rates and Figure 4.3 the distance errors. The hand precision rates are usually clearly lower than the other feature points precision rates. In “jack”, “wave1” and “wave2” actions the hands precision rates are better because the shape of the hands is less time within the interior of the silhouette contours. The reason why the right hand detection performs worse than the left on the “wave1” action is due to the fact that the action requires only the left hand doing the waving whilst the right is left along the torso, preventing its good detection. The good detection and matching of the hands is concluded to be higher when they are positioned well aside the torso during a longer period of time. In actions like “jump” or “skip”, most of the time the shapes of the hands are hidden inside the silhouette contours, which makes their detection harder to accomplish when merely using the silhouette contours.

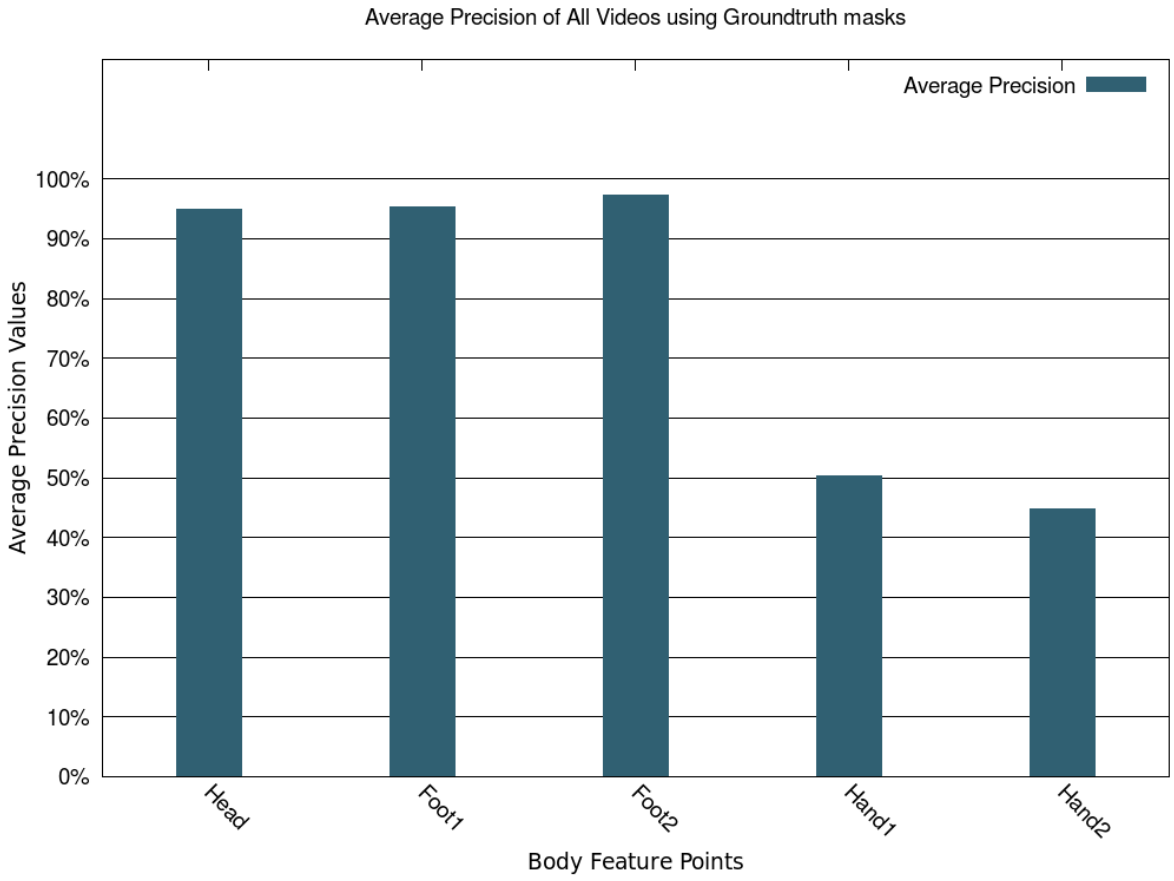


Figure 4.2: Average precision for all the videos on ground-truth masks

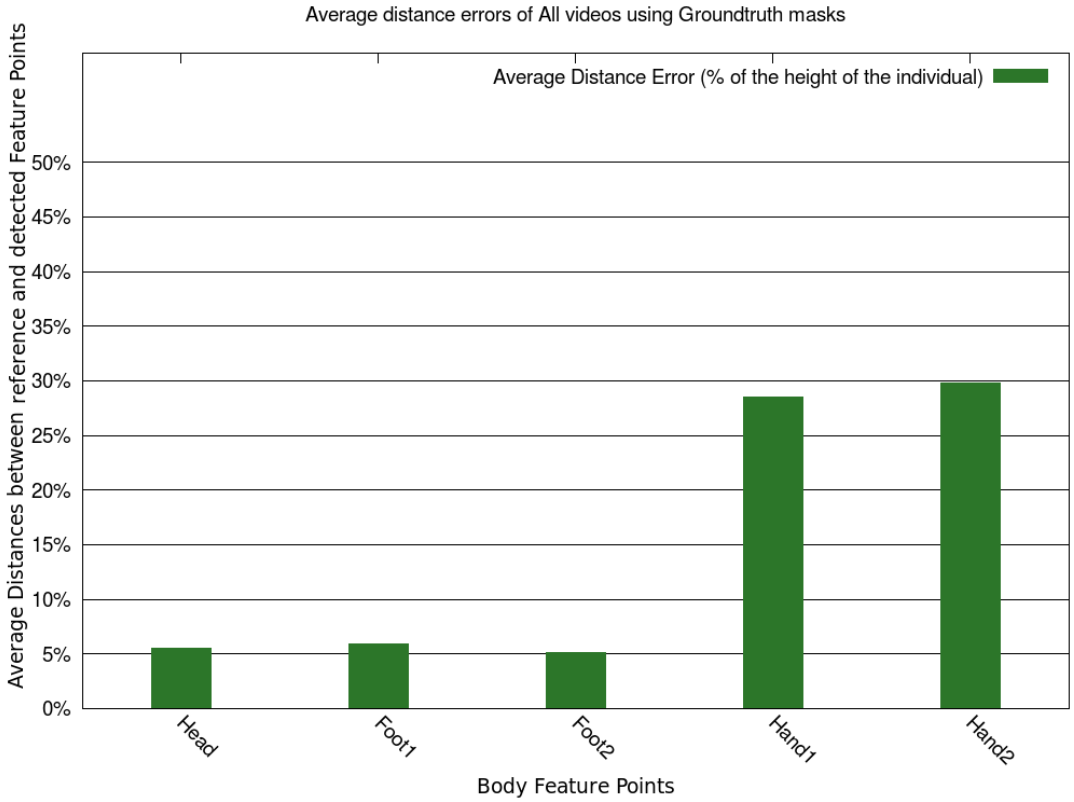


Figure 4.3: Average Euclidean distance errors for all the videos on ground-truth masks

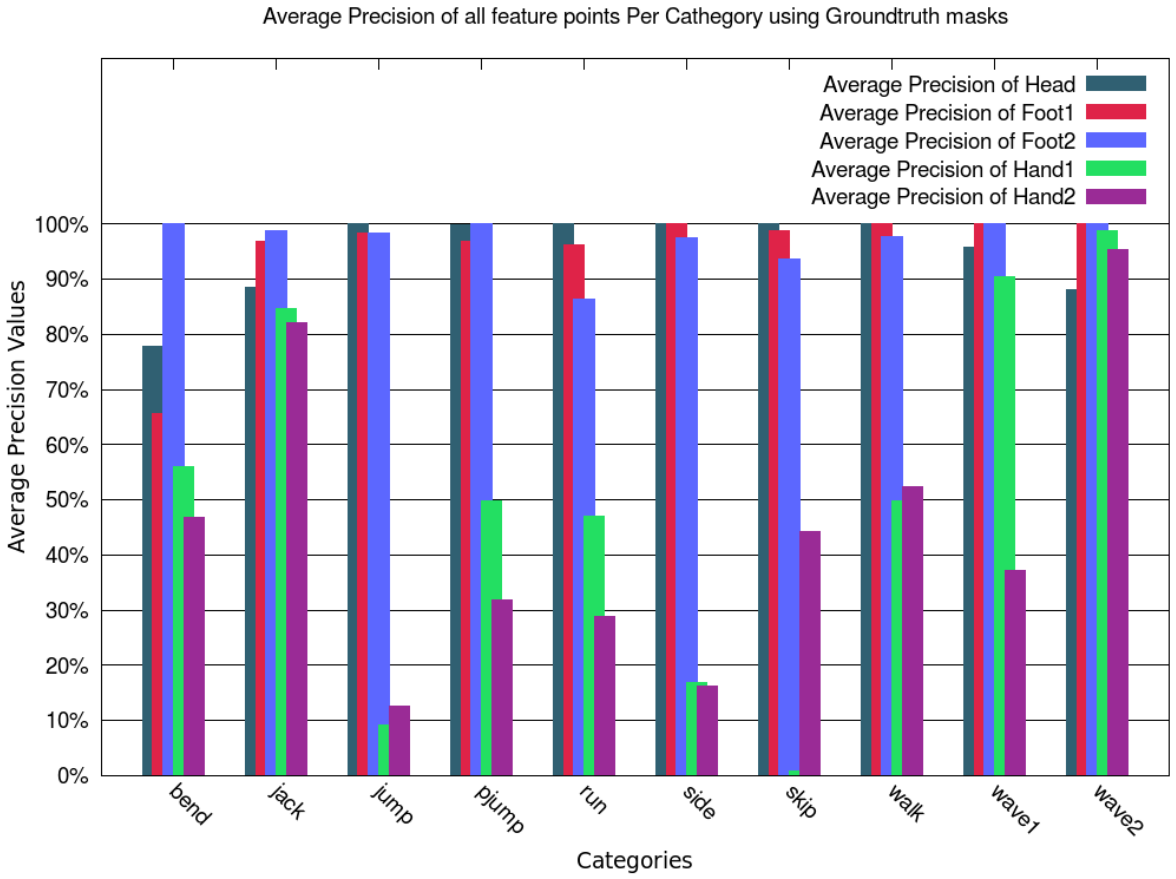


Figure 4.4: Average precision rates for each feature point on ground-truth masks

A good example of a pose where the hands do not stand out from the silhouette contours is illustrated in Figure 4.5. The right hand is so close to the torso that it is imperceptible in the extracted contours and in the respective combined distance plot represented in Figure 4.5(a) *i.e.* it does not originate a local maximum in it. Additionally, in the "wave1" action there is a considerable difference between the left and right hand detection. This is because the waving is with one hand only, being the other hand on the inside of the silhouette contours, thus not representing a local maximum in the combined distances contours and consequently poor detections occur.

In order to identify feet feature points it is necessary to firstly consider the area below the centroid to search for them (which always happened in all the videos) and take the left side of the centroid to look for the left foot and the right side for the right foot. Finally, the point from these sub-areas with the highest combined distance are considered the left and right feet feature points. Naturally, not always the feet are clearly on each side of the centroid. That is why whenever no local maximum (of the combined distance plot, further discussed in Section 3.2) is detected in one of these sub-areas, the corresponding coordinates of the foot are assigned to be the same of the other foot. Though this criteria may still induce some error, in cases where one both foot are on the same side of the centroid, and yet, quite separated, it still maintains low errors because this is an unlikely pose. If in fact both feet are on the same side of the centroid, they are likely to be close by or even side by side, for most poses.

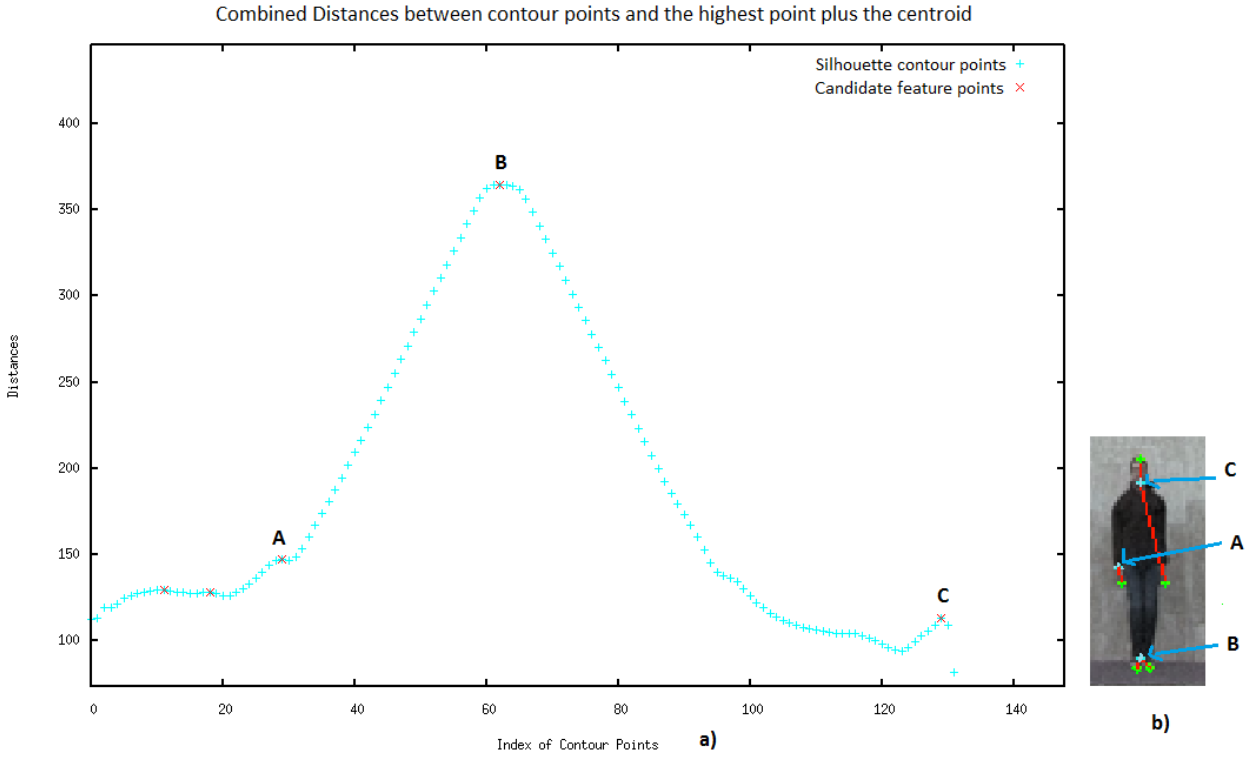


Figure 4.5: Example of bad hand detection using Background Subtraction; a) Combined distances plot with the identification of the detected feature points; b) Frame with reference and detected feature points in green and blue respectively. The red lines represent the distance error between them

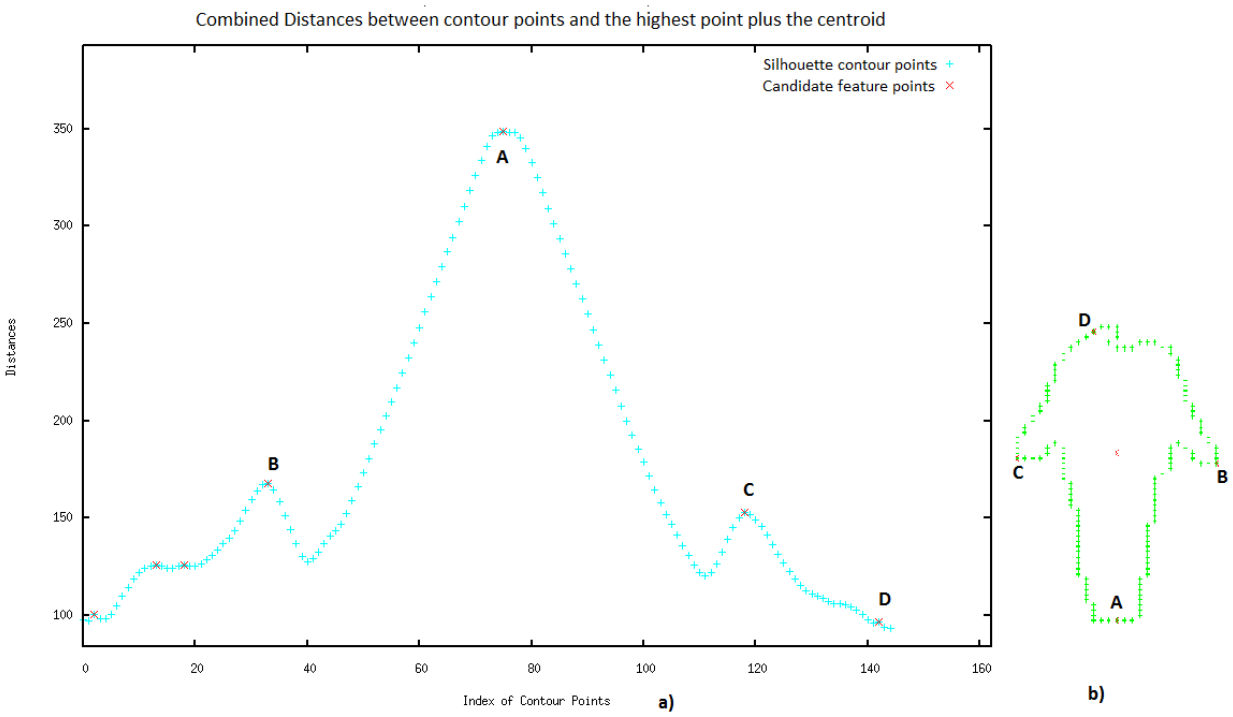


Figure 4.6: Example of good feet matching with two feet assigned to the same feature point; a) Combined distances plot; b) Extracted contours

In Figure 4.6 it is represented a good example of an effective application of this criteria. For instance in Figure 4.6(b) the extracted contours of a frame where both feet were assigned to the same feature point are represented. The matching of the detected feature points between the distances plot in Figure 4.6(a) is also performed. There is not a clear differentiation of both feet *i.e.* the legs are so close together that the foreground mask from which the contours are extracted from does not present any background pixels between them. As a consequence, there is only one local maximum below the centroid which is assigned as the feature point of one of the feet, depending if it is on the left or right side of the contours. Henceforth, there is not even one candidate feature point to consider for the other foot. As already addressed in Section 3.2, this case is resolved by assigning this foot the feature point of the other foot. This is confirmed with the exact coordinates assigned to each feature point for the same frame (Figure 4.6) presented in Table 4.4.

Feature Point (Point in Figure 4.6)	Euclidean distance between reference and detected points (in % of the height of the subject)
Head (D)	14%
Left Foot (A)	2%
Right Foot (A)	9%
Left Hand (C)	6%
Right Hand (B)	5%

Table 4.4: Detected and reference points coordinates and respective distance error for Figure 4.6

Moreover, another case in which at least one hand can be badly detected is when one or both hands reach above the head, which happens in the “jack”, “wave1” and “wave2” videos. However, particularly in “wave1” and “wave2” actions, this issue does not drastically affect the results because whenever the hands are above the head, the motion tend to get them closer to the head, which ultimately leads to the detected points assigned to the head and hands to be closer to each other. That said, if the feature points of the head and hands are close enough to be within the defined threshold used for precision calculation, no matching error would occur.

The combined distance plot is built based on the distance from the highest point of contours, which in the majority of poses is assumed to be the head, and from the centroid to all points of contours. The head is actually considered to be the nearest point to the highest point of contours, so if this point is for instance a hand, the further away it is from the head, the higher would be the distance error between the head and the detected feature point for it. An example of this situation is presented in Figure 4.7 where the right hand detection is actually affected in this case as well. This happens because it is considered to be the highest local maximum of the distance plot on the right side of the centroid above the feet. As it is visible in Figure 4.8(b) the head is located on the right side of the centroid, and it actually is a local maximum in the distance plot in Figure 4.8(a), hence its wrong detection as well as the right hand.

Figure 4.7: Frame illustrating a bad head detection using Ground-truth masks.The reference and detected feature points in green and blue respectively. The red lines represent the distance error between them (frame 14 of subject “Ido” performing “jack” action)

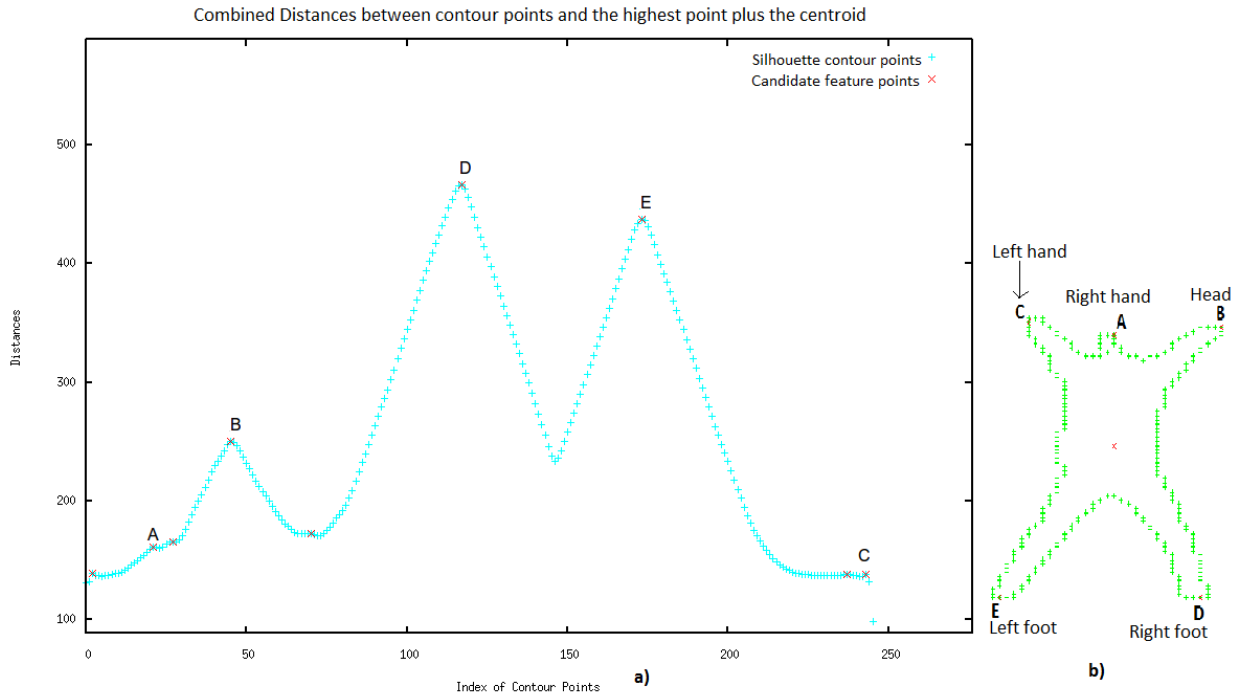


Figure 4.8: Combined distance plot and contours of Figure 4.7; a) Combined distances plot; b) Extracted contours with corresponding detected feature points

The "jack" action, which is the one being performed in Figure 4.7, is one of the few actions where there is an error of the head feature point because of this issue. The Euclidean distance errors are specially higher in the frames where the hands are positioned above the head, as it was possible to check in the distance to reference point over frames of the head feature point which is presents in Figure 4.9.

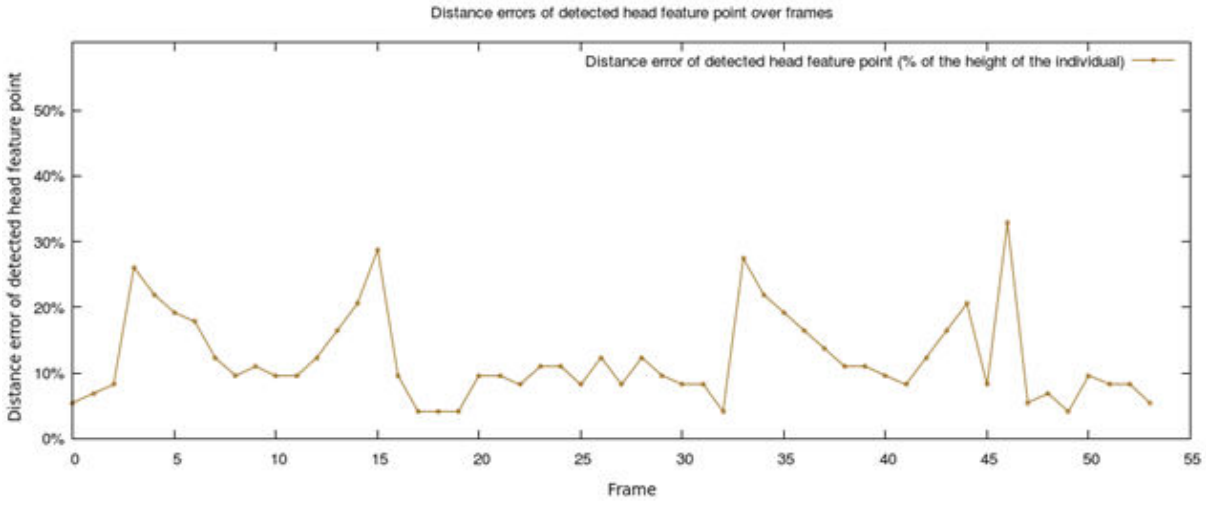


Figure 4.9: Euclidean distance to reference point of head in video of subject "Ido" performing "jack" action

A way to overcome this obstacle would be evaluating if the abscissa of the head feature point is nearby the centroid's abscissa, considering a threshold. For most common poses, the head is indeed aligned with the contours centroid.

Despite a few issues were identified, for the most common poses the feature points of the feet and head are correctly assigned. As for the hands, as results show, the detection is quite more

problematic. The proposed approach requires the limbs to stand out from the torso as much as possible. However, there are several poses where the outline of the hands is not perceptible from the perspective of the camera. Whenever the subject places the hands aligned with the torso, it becomes harder to detect them.

4.3 - Point detection evaluation

Even though the contribution of this work is the actual matching of the detected feature points and the part of the body to which it corresponds to, an interesting perspective of evaluation is to evaluate to what extend the correct points are detected but badly matched. Nevertheless, it is important to keep in mind that the results presented with the detection-only version of the algorithm take in consideration the ground-truth videos and the distance to nearest-point. On the other side, the verified improvements may be seen as how much the results can get better by processing the five detected feature points without assigning new values based on alternative procedures to any of them. In addition, this allows comparing results with previous work that focus only on the detection of the points.

In Figure 4.10 the average precision per category of the ground-truth videos considering the distance to reference point is confronted with the corresponding precision of the ground-truth videos considering the distance to nearest point. It should be noted that the precision rate when using the distance to nearest point (detection-only version of the algorithm) is an indicative value. It is not a realistic measure when it is used the distance to nearest point because this distance is calculated based on the reference feature points, as already mentioned. Thus, it should be analyzed as a mean to compare the magnitude of the change in the results when using the distance to reference point and the distance to nearest point.

The actions that present a higher improvement are “bend”, “run”, “skip”. This means that in the videos of those actions it often occurred that for each of the 5 detected feature point there was at least another detected feature point that was nearer of the corresponding reference point.

The feature points of the hands are the ones whose difference between the distance to nearest point and the distance to reference point is bigger, as Figure 4.11 shows. It also shows that the average distance to nearest point for both feet of the algorithm is lower than the distance to the reference point. On several foot detection errors, both feet are nearby each other and at least one of them is detected correctly, much like the situation addressed in Figure 4.12.

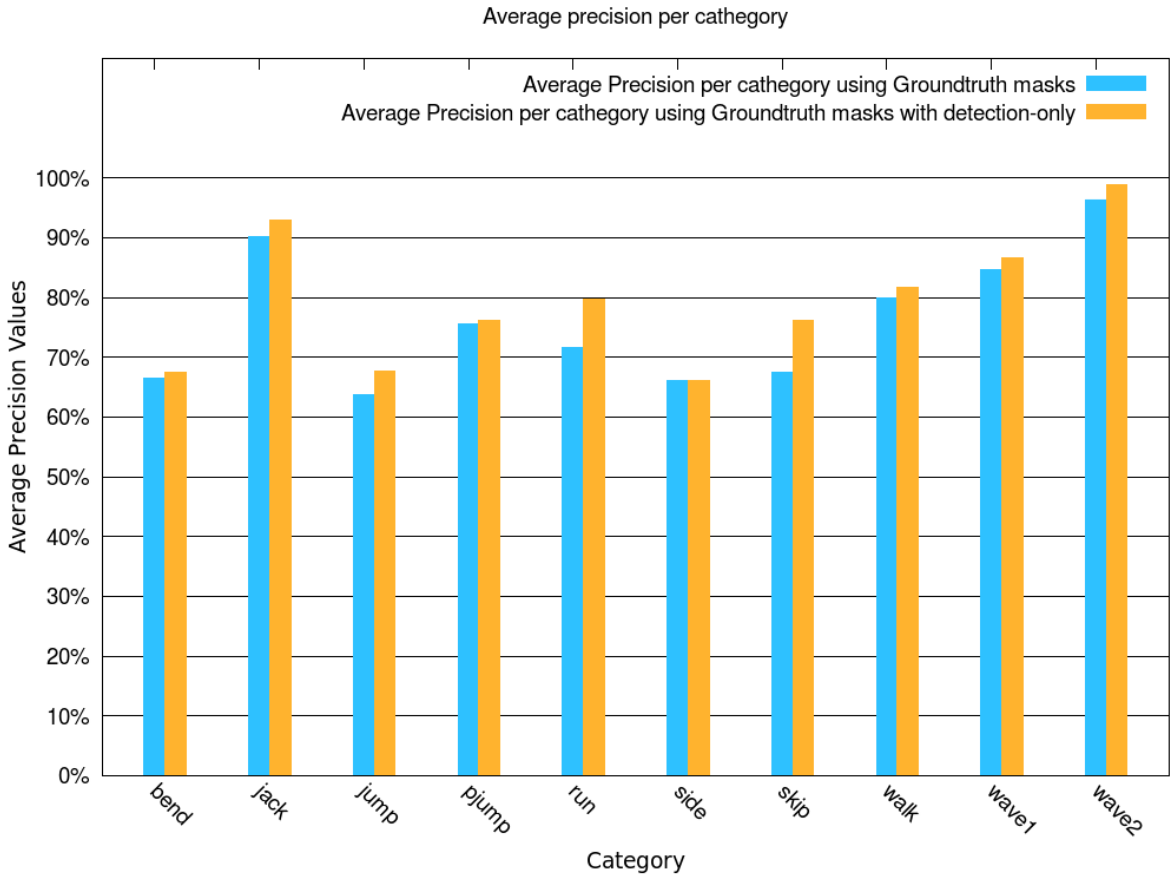


Figure 4.10: Average precision per category using Ground-truth masks with and without the detection-only version of the algorithm

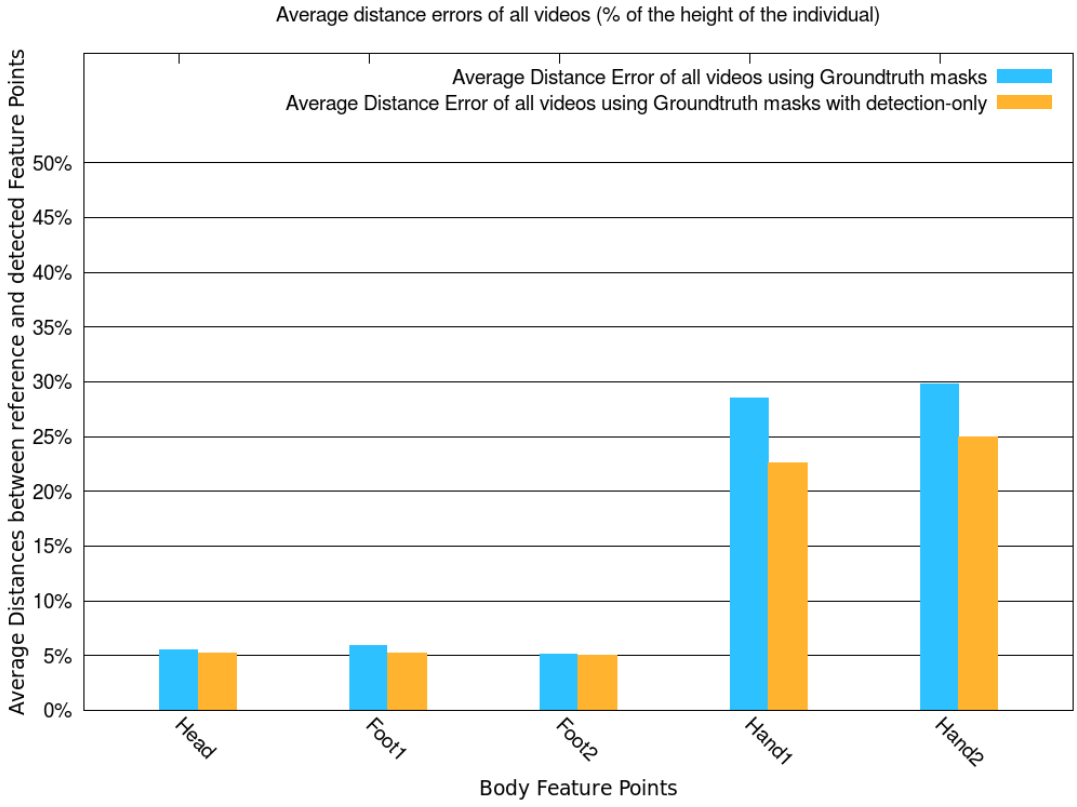


Figure 4.11: Average distance to reference point and to nearest point using Ground-truth masks

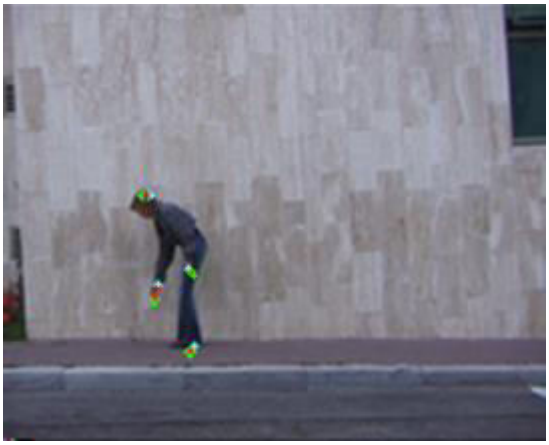


Figure 4.12: Example of good matching with distance to nearest point. The reference and detected feature points in green and blue respectively. The red lines represent the distance error between them

The feature points of the hands are indeed the points whose detection presents a higher degree of uncertainty, and whenever they are both positioned on the inside of the silhouette (which happens during the "skip" and "run" actions), they are most likely close to each other. This means that if a feature point is correctly assigned to one of the hands, the detection-only version of the algorithm assigns the same point to both hands, eliminating the error in several cases. It is possible to verify this scenario by analyzing Figures 4.14 which show the distance errors of the left hand of a video belonging to the "run" action in the normal version of the algorithm in ground-truth videos and the same distance errors in the detection-only version of the algorithm in ground-truth videos. The error is lessened on the latter because the right hand presents less errors and its feature point is often considered for the left hand as well, as it is visible in Figure 4.13. Obviously this results in better matching precision rates, however they are not realistic, since the detection-only version of the algorithm re-computes the 5 feature points taking in consideration the knowledge of the 5 reference feature points.

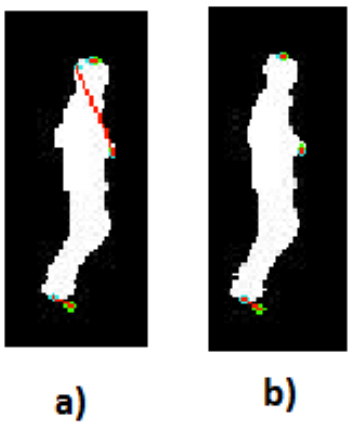


Figure 4.13: Example of elimination of hand error with the detection-only version of the algorithmThe reference and detected feature points in green and blue respectively. The red lines represent the distance error between them a) Frame of a video of the action "run" processed with the regular algorithm; b) Same frame as in Figure 4.13 (a) processed with the detection-only version of the algorithm

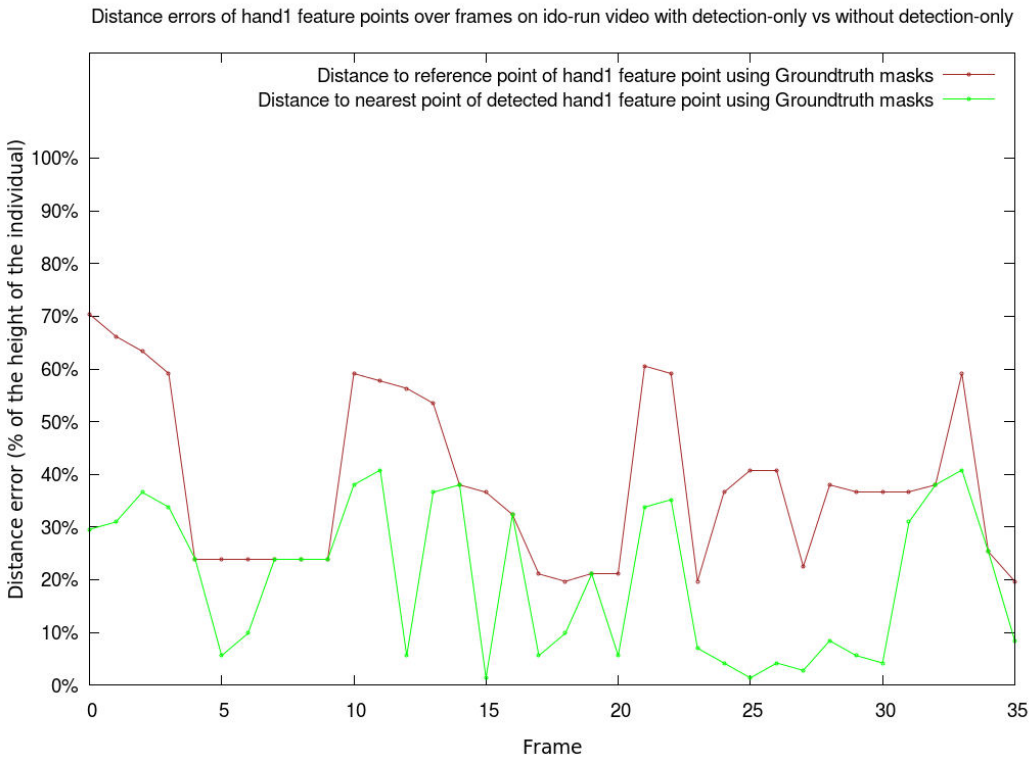


Figure 4.14: Distance to reference point and to nearest point of left hand of "run" action video using Ground-truth videos

The results presented in this Section are the most optimized from all the experiments performed. Some matching errors addressed in Section 4.2 are resolved using this evaluation method. However, it is important to remember that the detection-only version of the algorithm considers the reference feature points to recalculate the detected ones. Obviously this is not a valid way of obtaining the matching results. It can only be seen as a mean to evaluate the detection of the feature points, independently of the part of the body they would be assigned to by the algorithm.

4.4 - Impact of post-processing on matching

In this section the post-processing of the algorithm is evaluated comparably to its absence. The average distance errors in Figure 4.15 do not present noteworthy alterations. The detection and subsequent matching of the five feature points did not suffer substantial alterations with the post-processing, which is verified by the average precision rates per category in Figure 4.17. This leads to conclude that the criteria chosen to combine the Freeman chain-code algorithm feature points with the 5 feature points outputted by the proposed algorithm did not substantially influenced their according re-computing.

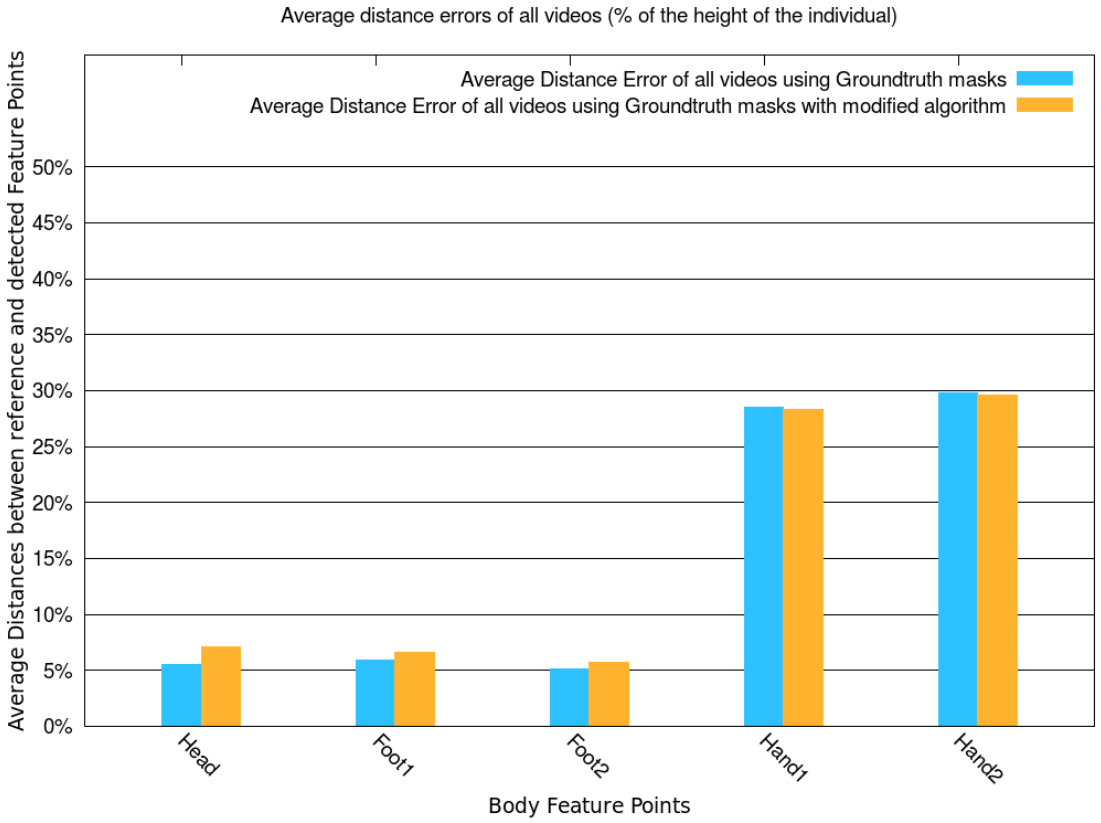


Figure 4.15: Average distance to reference point with and without post-processing using Ground-truth masks

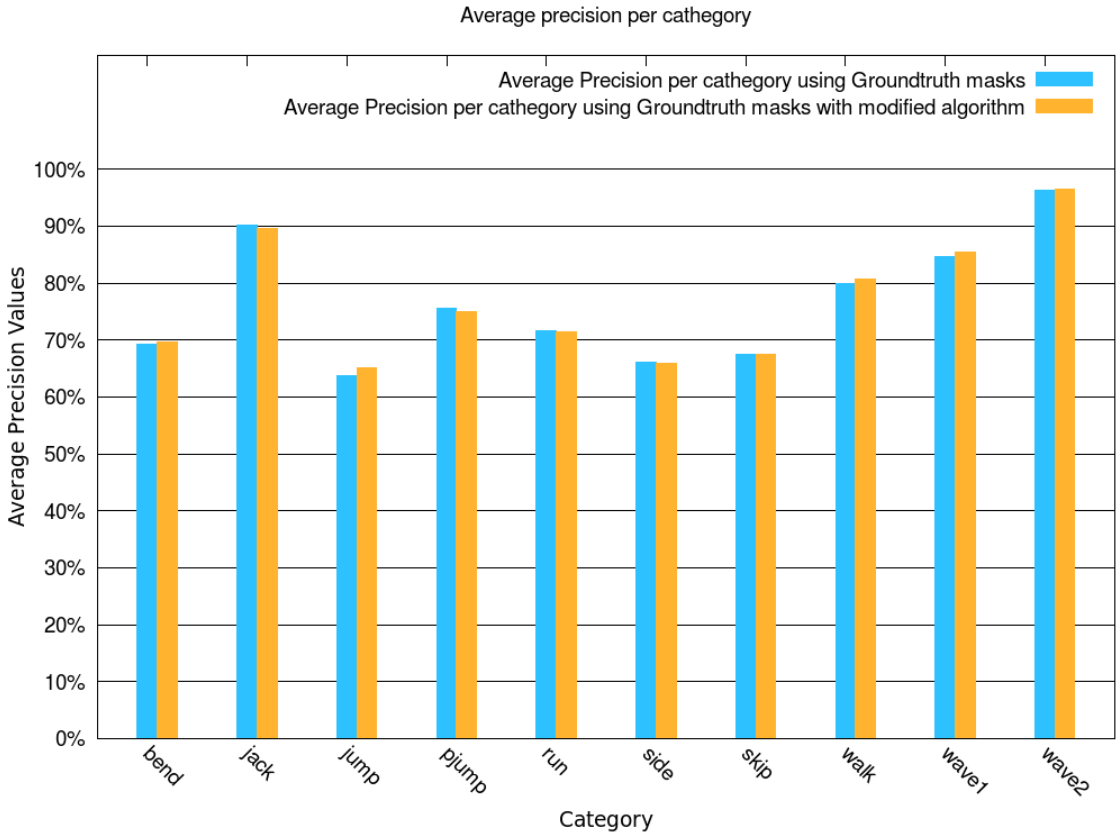


Figure 4.16: Average precision values per category with and without post-processing using Ground-truth masks

In some situations the post-processing did improved the detection distance error *i.e.* it recalculated the feature point making it nearer its reference location. In Figure 4.17(a) where the post-processing took place, the distance error of the head feature point is slightly lower, as the smaller red line uniting the reference and detected point indicates and the right hand feature point is recalculated from the left side of the centroid to the right. In fact, in Figure 4.17(b) the same feature point (even though it is on the left side of the centroid) is assigned to both hands, which indicates that on the right side of the centroid, no local maximum of the combined distance plot was detected. Henceforth, the post-processing permitted a better feature point recalculation in this particular case.

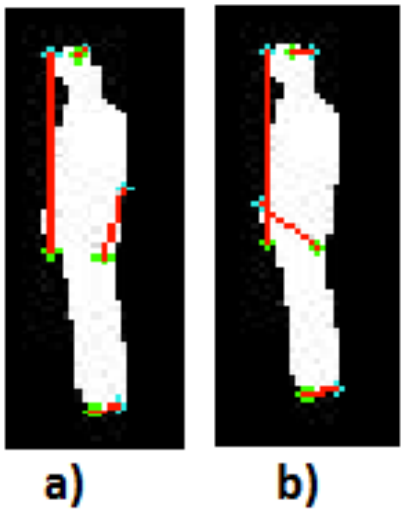


Figure 4.17: Example of detection improvement originated by the post-processing using Ground-truth masks. The reference and detected feature points in green and blue respectively. The red lines represent the distance error between them a) Frame with post-processing b) Same frame of Figure 4.17(a) without post-processing

The adopted post-processing method has the potential of being improved. The simplistic approach of considering the nearest feature point of the Freeman based method (presented in Section 3.3) does not consistently permit the achievement of better detections of the proposed algorithm.

4.5 - Ground-truth vs masks obtained with background subtraction

In order to test the algorithm performance in perfect conditions, the silhouette of the subject to be treated would have to be extracted in an exact way in order to obtain the corresponding contour points accurately. However, that is a constraint we have to take in consideration on the analysis of the results of this experiment. In this section, the errors derived from a defective background subtraction are identified and its impact on the overall results are evaluated.

Figures 4.18-4.19 show the average precision, Euclidean distance and angle errors for all the 30 videos respectively, with the usage of background subtraction and the ground-truth videos as inputs. As it is possible to see in Figure 4.18, it is clear that the head and feet have a higher score

when using the ground-truth videos, which leads to think that the background subtraction indeed cause relevant noise.

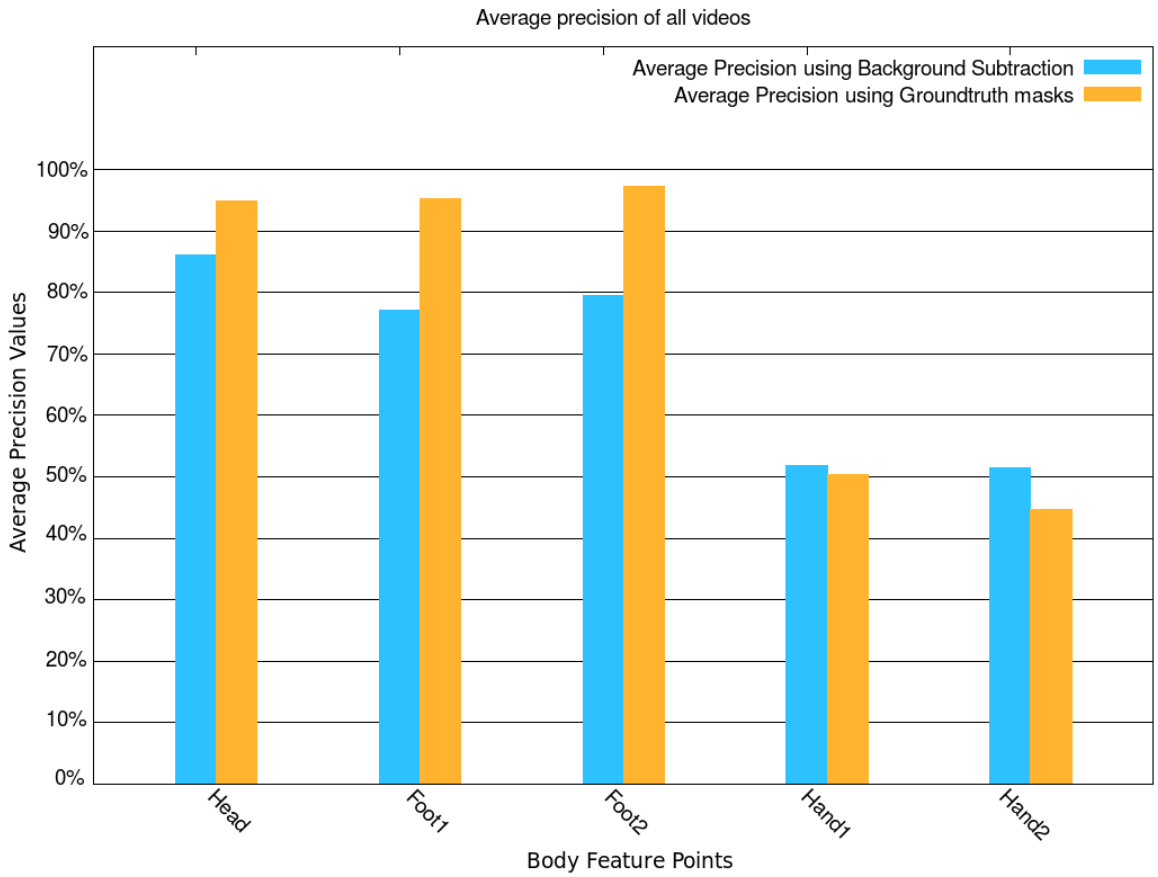


Figure 4.18: Average precision using Background Subtraction and Ground-truth masks

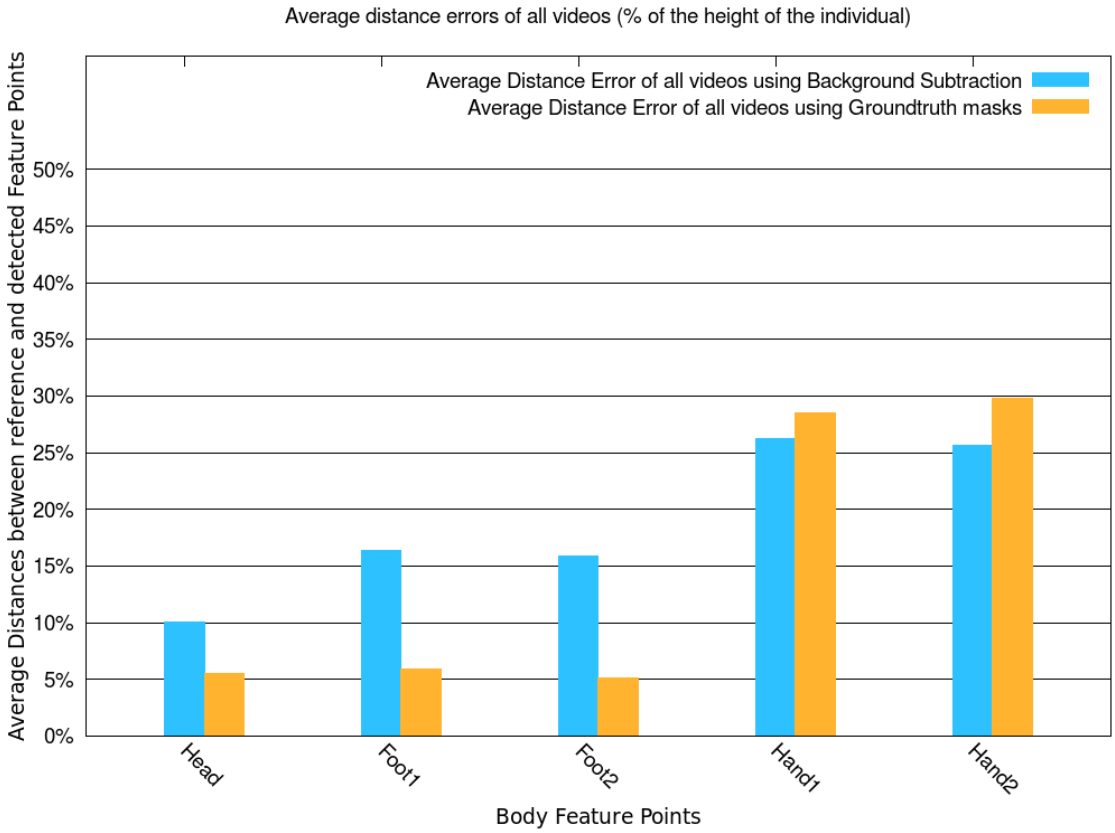


Figure 4.19: Average distance to reference point error using Background Subtraction and Ground-truth masks

However the precision of the hands is higher when using background subtraction, which may seem incongruous in a first analysis, but taking in consideration the fact that for the great majority of the videos the hands are close to the body, the contours become homogeneous hence preventing a good detection of them by simply having the silhouette contours data. Either way, the distance errors in Figure 4.19 show according results in the same feature points, the errors are higher for the feet when using background subtraction and the opposite for the hands. In order to further understand why the hands detection is actually better when using background subtraction, it becomes necessary to take a closer look at the results by examining same extracted contours.

On the other hand, one aspect that characterizes background subtraction is the noise that it implies to the contours. Considering the way hands are detected, by selecting the two contour points above the highest foot with the highest combined distance on each side of the centroid (see Figure 3.4(b) and Section 3.2), if the extracted contours don't have many irregularities, which is often the case of the ground-truth masks, it is not likely that the plot of combined distances presents many local maximums. Thus, the degree of uncertainty increases, especially in the case where hands are along the torso. For example, in Figure 4.20(c) and Figure 4.20(d) the extracted contours obtained with background subtraction and from the ground-truth image of the same frame of a video containing a subject running are represented alongside the respective combined distances plots in Figure 4.20(a) and Figure 4.20(b) respectively. It is noted in Figure 4.20(b) that there exists less local maximums because the extracted contours are more homogeneous. In some

cases, this may mean better detections of the hands. In fact, that was the case in this video, as it is possible to verify in Figure 4.21 where the average distance errors corresponding to the video of Figure 4.20. It is visible a slightly higher error on both hands feature points for the ground-truth frames.

By evaluating the results separately for each category of action, different conclusions can be discussed. Figure 4.22 represents an average of the precision values of all the five feature points for both the background subtraction and the ground-truth videos for each category. Figure 4.23 shows the average precision per category for each feature point using Background Subtraction. The head and both feet present the better matching rates for the majority of videos, as it happened with the Ground-truth masks, but with a generally lower precision rate than the latter.

Moreover, it was detected major flaws in the videos of one specific subject in different actions *i.e.* the background subtraction was performing particularly defectively in this case. This reflected in poor results, which were notable more plainly in the “wave1”, “wave2” and “jack” videos in every feature point, thus in these cases the overall performance of ground-truth images is particularly better, as it is visible in Figure 4.22. Nevertheless, on the exception of a few cases, the general tendency is to obtain better results with the ground-truth images. This tendency is even more obvious when it comes to precision values of feet and the head feature points per category.

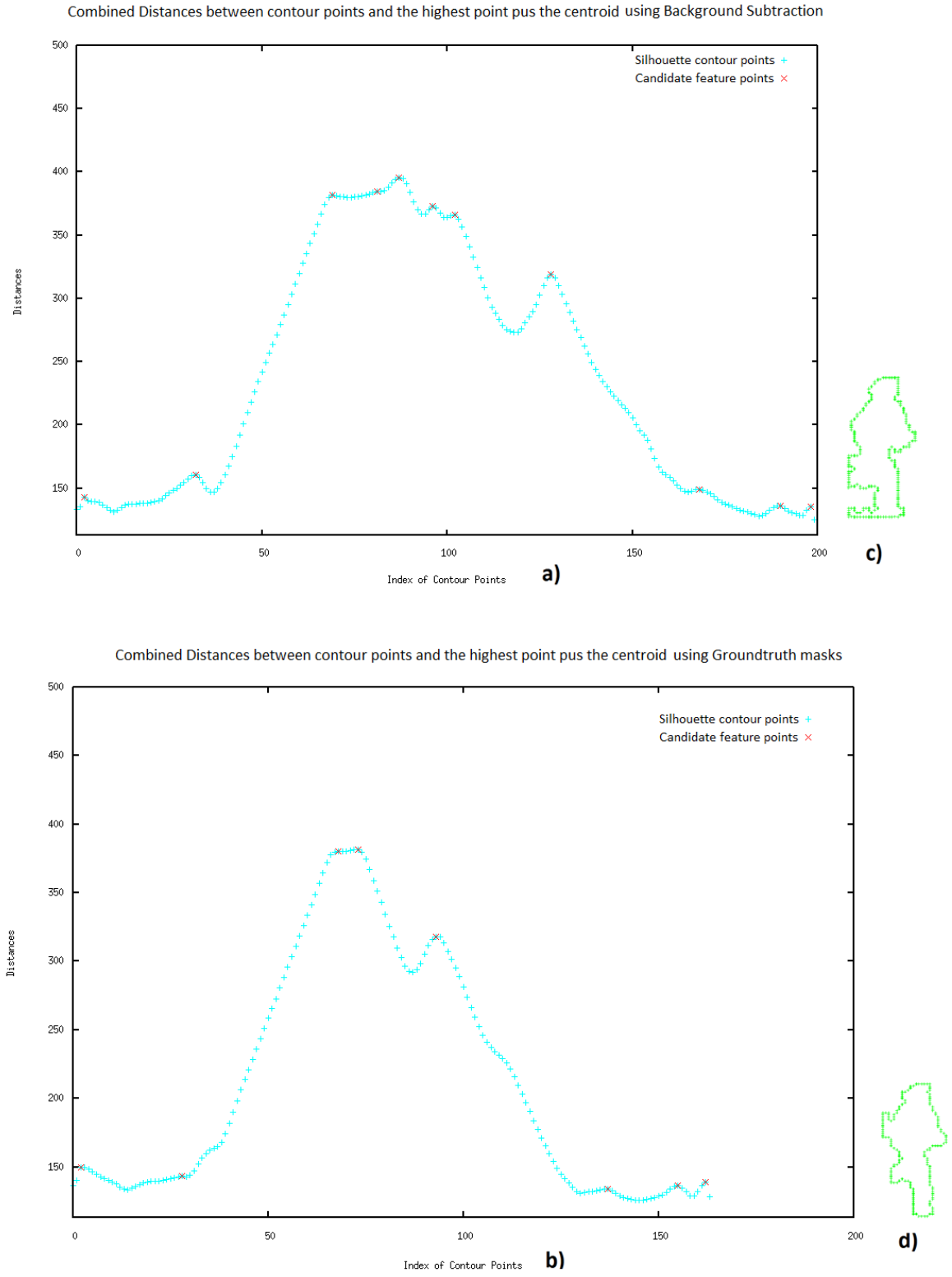


Figure 4.20: Combined distance plots and corresponding contours for background subtraction and ground-truth mask; a) Combined distance plot for background subtraction; b) Combined distance plot for ground-truth mask; c) Extracted contours for the background subtraction; d) Extracted contours for the respective ground-truth mask

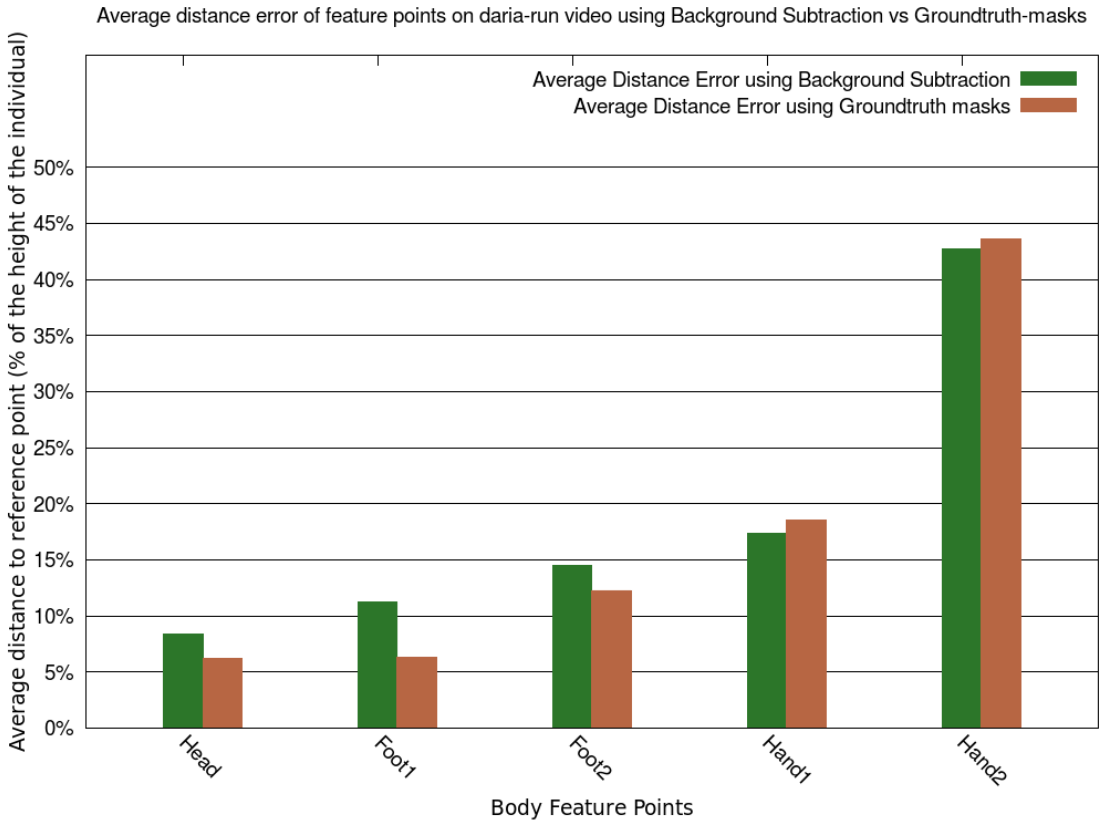


Figure 4.21: Average distance to reference point for video of frame represented in Figure 4.20

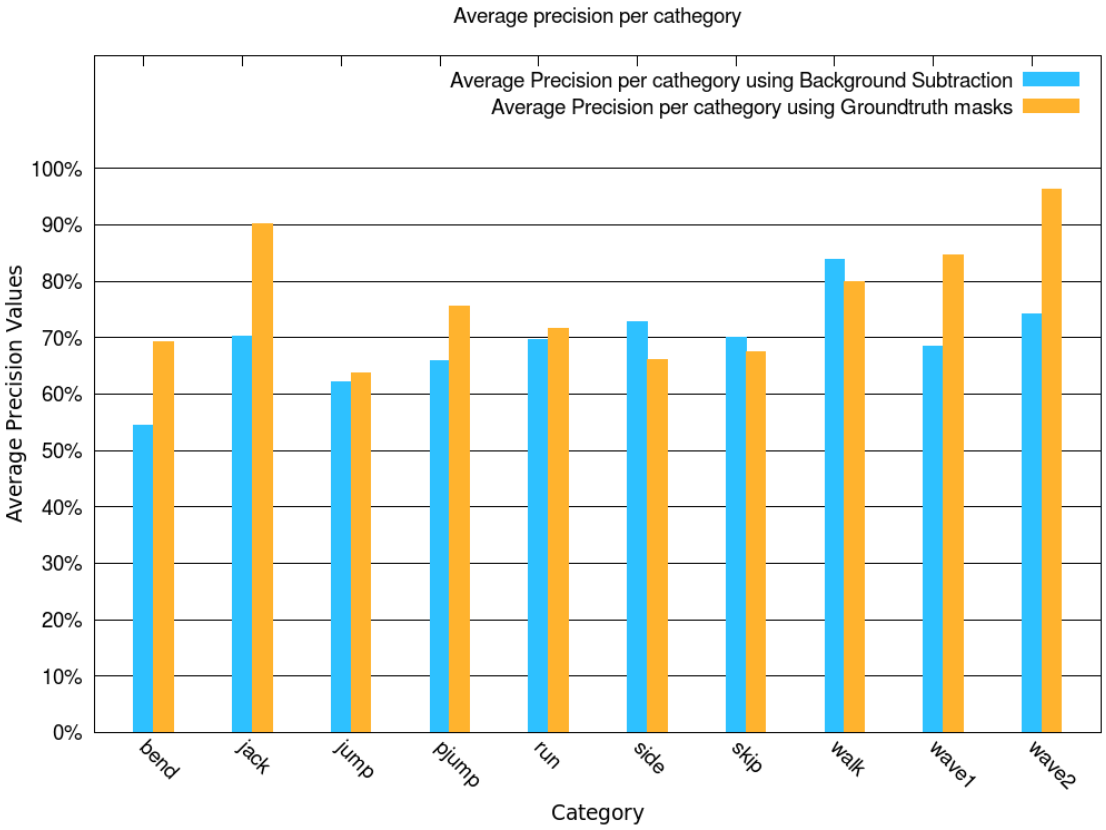


Figure 4.22: Average precision per category using Background Subtraction and Ground-truth masks

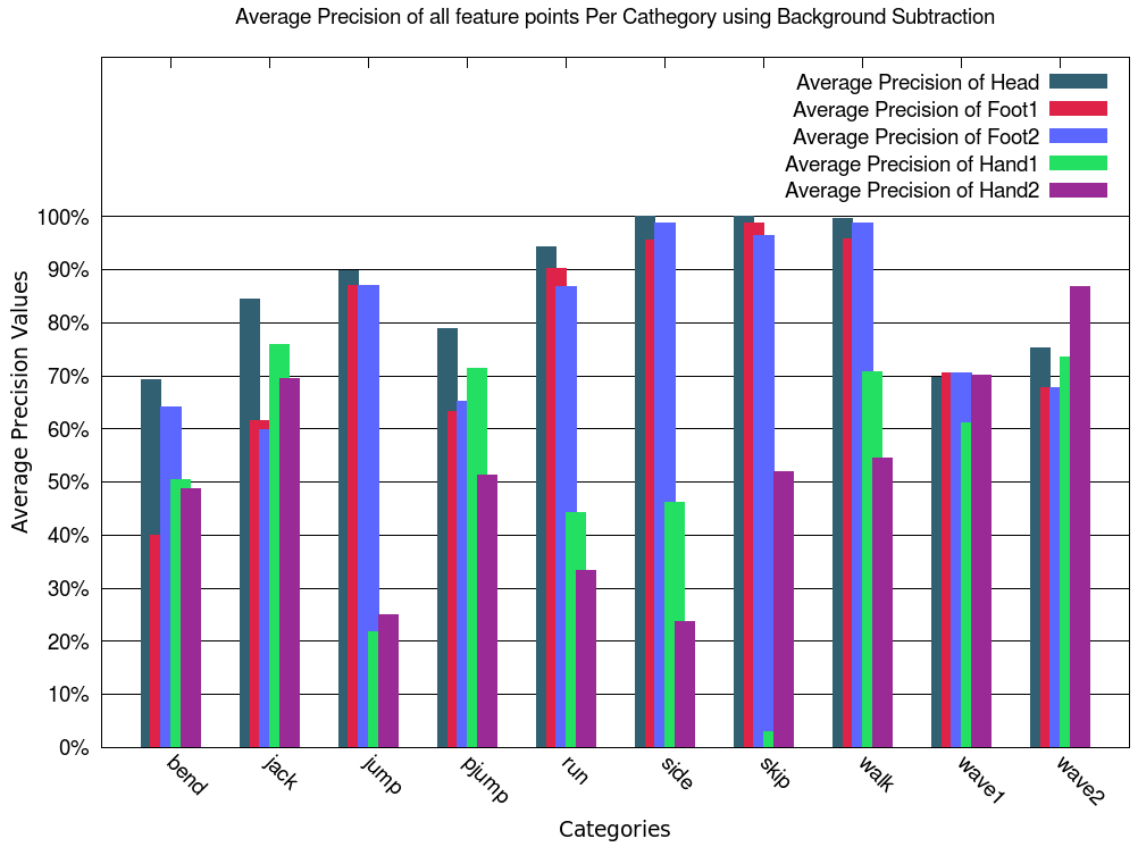


Figure 4.23: Average precision for each feature point per category using Background Subtraction

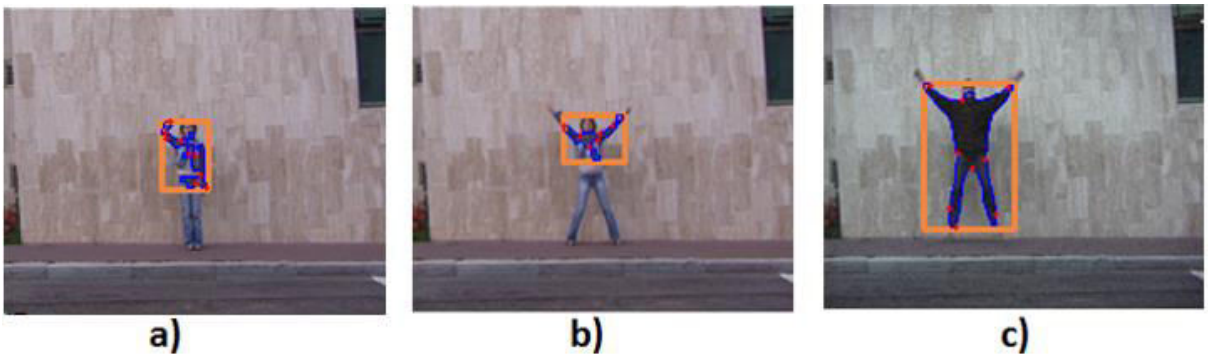


Figure 4.24: Examples of background subtraction errors; All the candidate feature points are represented with red dots and the contours with a blue line. The orange square covers the extracted contours. a) Major background subtraction error in feet on "wave1" video frame; b) Major background subtraction error in feet on "jack" video frame; c) Slight background subtraction error on hands on "jack" video.

By examining some examples, it becomes clearer the limitations imposed by the background subtraction. For example, noting that in Figure 4.24, the orange rectangle surrounds all the detected contours, it is clear that it Figures 4.24(a-b) the background subtraction performed quite badly. Particularly in Figure 4.24(a), it is one of the cases where a fixed background is used, the fault could be attributed to the misleading background image. However, besides that it was confirmed that it was not the case, in Figure 4.24(b) where the same subject is performing the action named "jack", where a MOG background subtractor is used, similar defects are verified. On the other hand, a different subject performing the same action of Figure 4.24(b) is illustrated in

Figure 4.24(c), but then again with a different background contrast. In this case, the dark clothes allied with the light background are a key factor for a good performance of the background subtraction process. The subject in Figures 4.24(a-b) wears lighter clothes and considering the different background contrast, it blends in it.

Occasional background subtraction errors were also identified by analysing discrepancies in the distance errors of the feature points over the frames. One example can be addressed by examining Figure 4.25 which correspond to the distance errors on the left and right foot, on a video of "jack" action. For instance, in frames 45 there is a peak on the distance error of the detected feature points of both feet. By taking a look to the matching frame in Figure 4.26(a), the background subtraction detected the legs only halfway, by the zone of the femur. On frame 48, represented in Figure 4.26(b), even though the subject is in a similar position, that error disappears. On another error peak, this time only on the right foot, between frames 13 and 18, by looking at the extracted contours on one frame in this interval, it is clear to see that the background subtractor did not detect the whole right leg, hence an increase on the detection error.

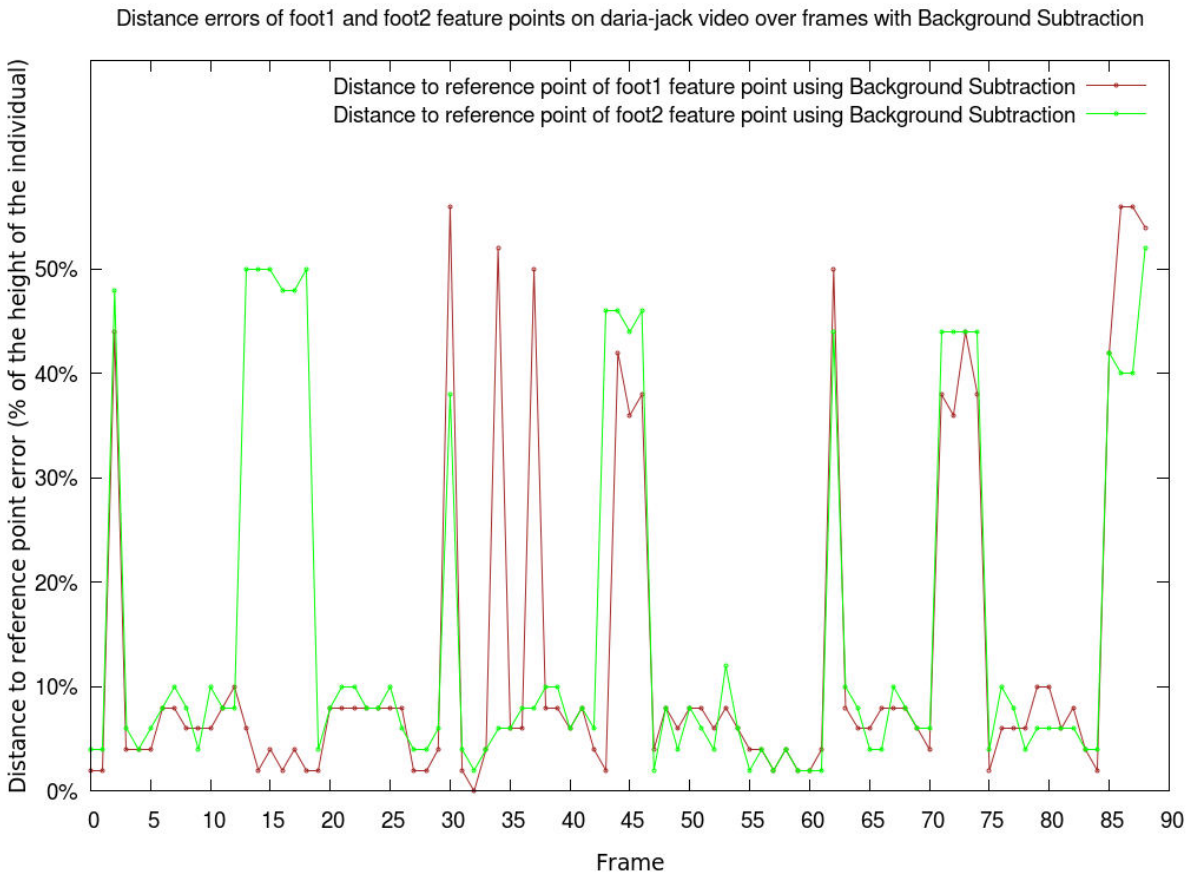


Figure 4.25: Distance to reference point of left and right foot over frames using Background Subtraction on "jack" action video

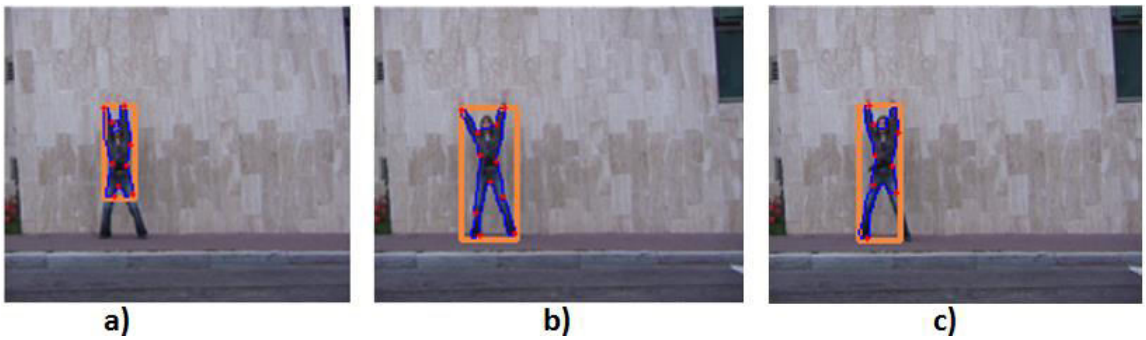


Figure 4.26: Background subtraction errors on the feet; All the candidate feature points are represented with red dots and the contours with a blue line. The orange square covers the extracted contours. a) frame 45 of "jack" action of subject "daria"; b) frame 48 of "jack" action of subject "daria"; c) frame 15 of "jack" action of subject "daria";

A poor foreground mask obtained by background subtraction was indeed in many cases an important factor for detection errors. Nonetheless whenever the conditions were reference for its proper functioning *i.e.* there was a good contrast between the subject and the background, it did produced acceptable results. However, in this experiment, there were actions which did not implicated a constant complete body motion, thus requiring a fixed background that was considered. In real life conditions, this may be hard to accomplish, especially considering lightning conditions and highly moving background scenes. This stresses the need for alternative ways of performing the foreground segmentation without undermining a good silhouette contours extraction.

4.6 - Evaluation of shape-related measures

In this Section the results are evaluated with particular focus on the proposed angles. For the purposes of this work, they are studied from an error perspective and additionally, they are also addressed as an anatomic measure that can provide characteristic patterns for each type of movement.

In Figure 4.27 the average angle errors for all the videos while using the ground-truth masks are presented. Both hands present a higher error than the feet. On the other hand, the angles *β3* and *β4* which are associated with the hands present a higher error. In Figure 4.28 it is possible to see that the actions that present a lower error on the two angles referred above are "jack", "wave1" and "wave2", which are also the actions with a lower distance to reference point error in both hands, as discussed in Section 4.2.

However, a greater distance error may not necessarily mean a greater angle error for that matter. It may happen that the angle error is zero and the distance error may be of notice. Figure 4.29 illustrates a situation where the four angles are exactly the same for the detected and reference feature points. However, there is a distance error, since the detected points are not in the precise same position as the reference feature points. Of course this is an extreme example, though it exemplifies how the distance errors may present no relation to the angle errors. Even so,

if a distance error takes place, the angle error should increase in the same proportion of the distance between the detected feature point and the:

- Straight line CF in the case of the angle $\beta 1$;
- Straight line CD in the case of the angle $\beta 2$;
- Straight line HG in the case of the angle $\beta 3$;
- Straight line HE in the case of the angle $\beta 4$.

On the other hand the measured angles present characteristic patterns for different actions. For instance, the reference right foot angles for a "run" action video for three different subjects are presented in Figure 4.30. Even though the mean angle appears to be higher in Figure 4.30, it does present similar periods where the angle continually increases and then decreases, much like a sinusoidal wave with an associated period. Though, the pace at which the subjects are running, the width of each step and the length of the leg directly influences the mean angle, the amplitude peak to peak and the frequency. Indeed, the subject "ido" is taller and has longer legs and additionally, he takes longer steps while running, thus the period of the approximated sinusoidal wave in Figure 4.30 corresponding to him appears to be higher than the other subjects. Despite these are the reference angles of the corresponding videos, the average angle errors for both feet in each of the two subjects for this specific action was near 5º, hence it is not likely that the overall behaviour of the detected angles over the frames would be substantially altered.

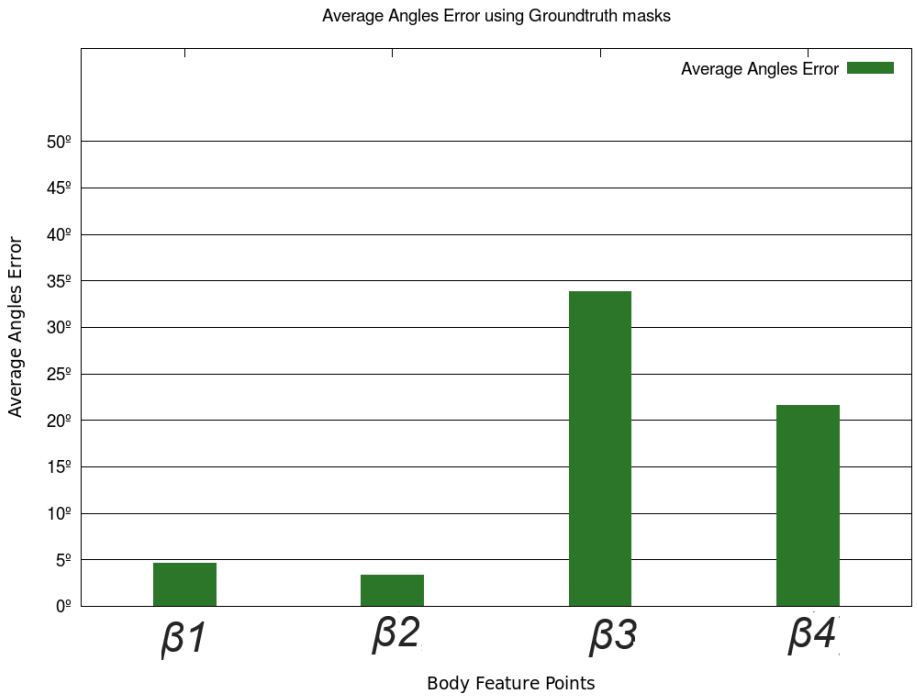


Figure 4.27: Average angle errors for all videos using Ground-truth masks (Bars from left to right *β1, β2, β3, β4*)

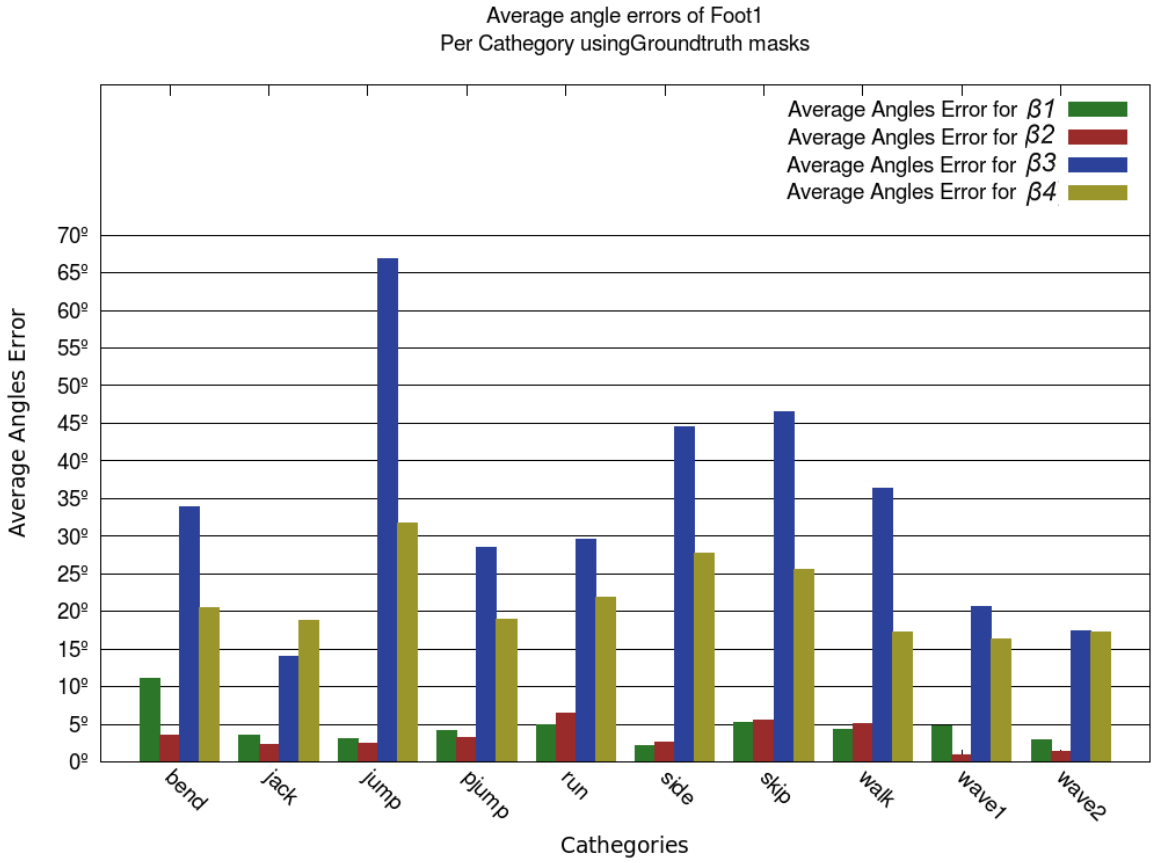


Figure 4.28: Average angle errors per category using Ground-truth masks

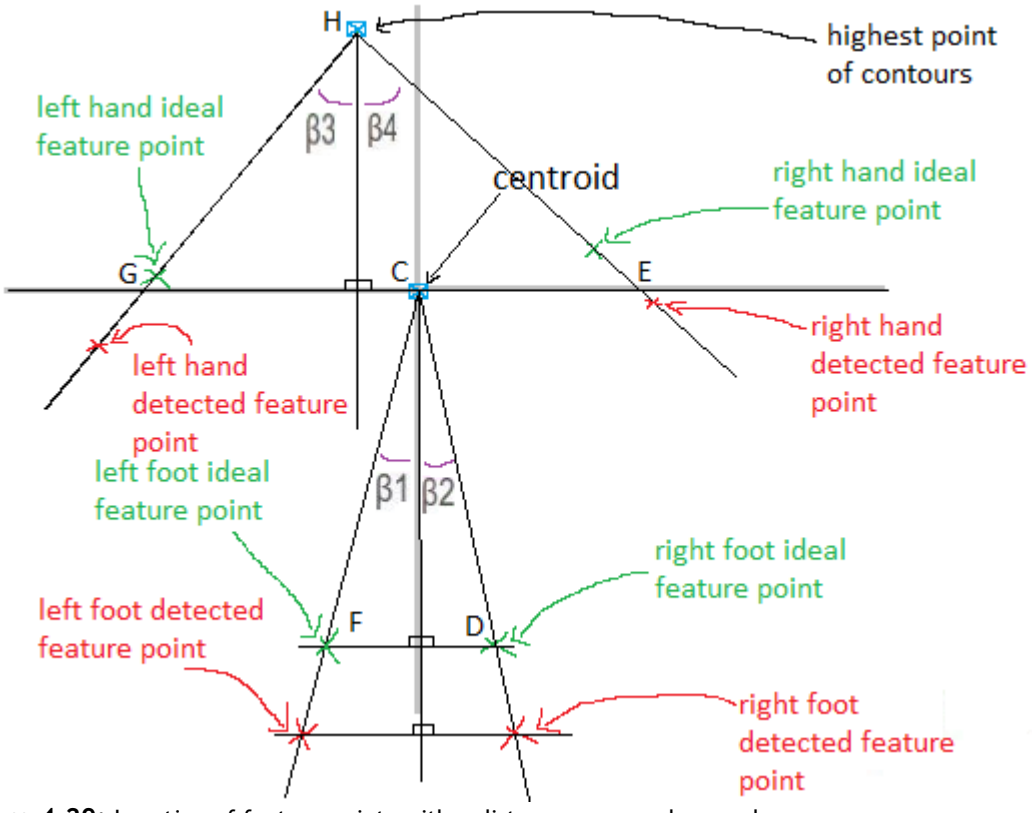


Figure 4.29: Location of feature points with a distance error and no angle error

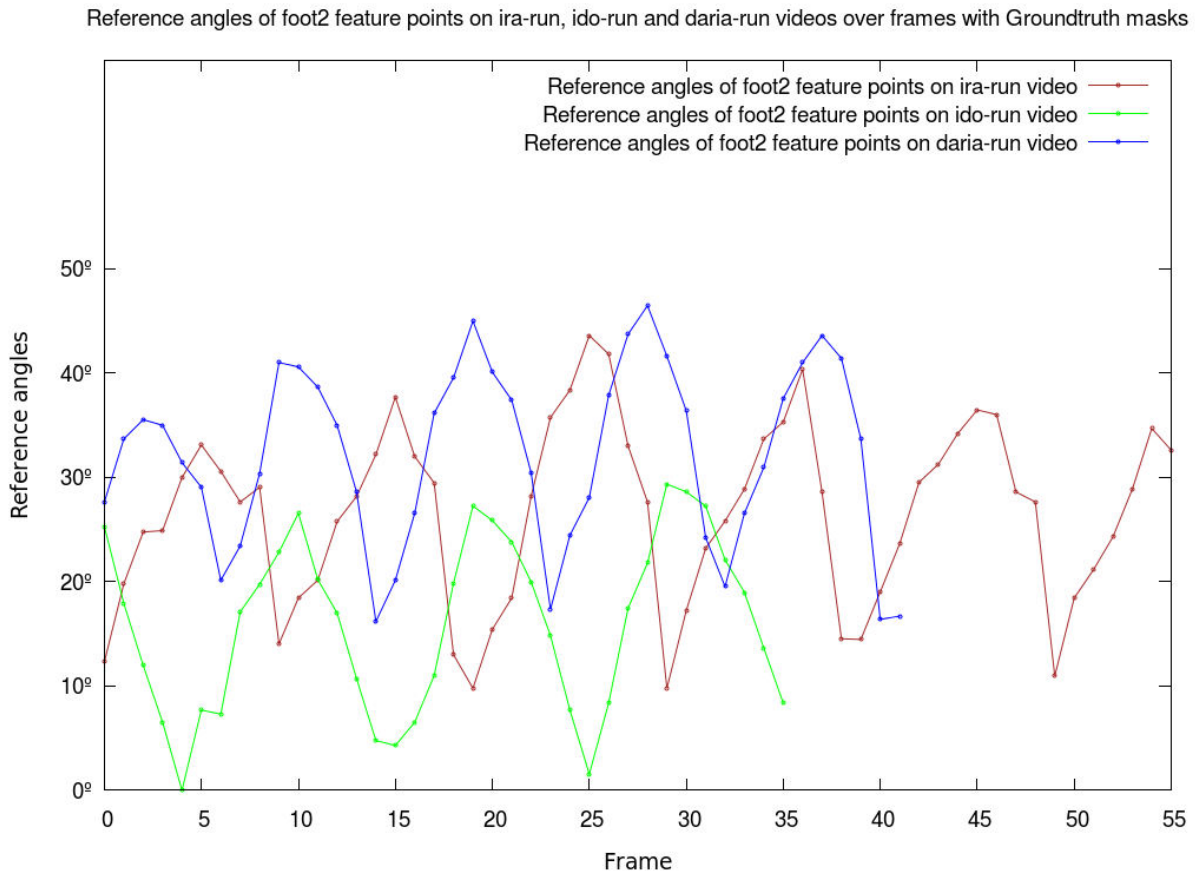


Figure 4.30: Right foot reference angles for the "run" action video of subjects "ira", "daria" and "ido" using Ground-truth masks

On the "wave1" action, where the subject waves with the left hand while standing still, a pattern is detected as well on the angles of the left hand. The left hand reference angles of subjects "ido", "ira" and "daria" of "wave1" action are presented in Figure 4.32. In the ascendant movement of the left arm, the corresponding angle increases until its maximum value which would be 90º if the feature point acquires the exact x-coordinate of the centroid (Figure 4.31(b)). On the descendant movement, the angle decreases until the arm reaches its standing position (Figure 4.31(a)). It should be noted that the subjects are not performing the motion at the same pace, hence in Figure 4.32 the referred angle pattern has different frequencies for each one of them.

Figure 4.31: Poses of "wave1" action of subject "ido"

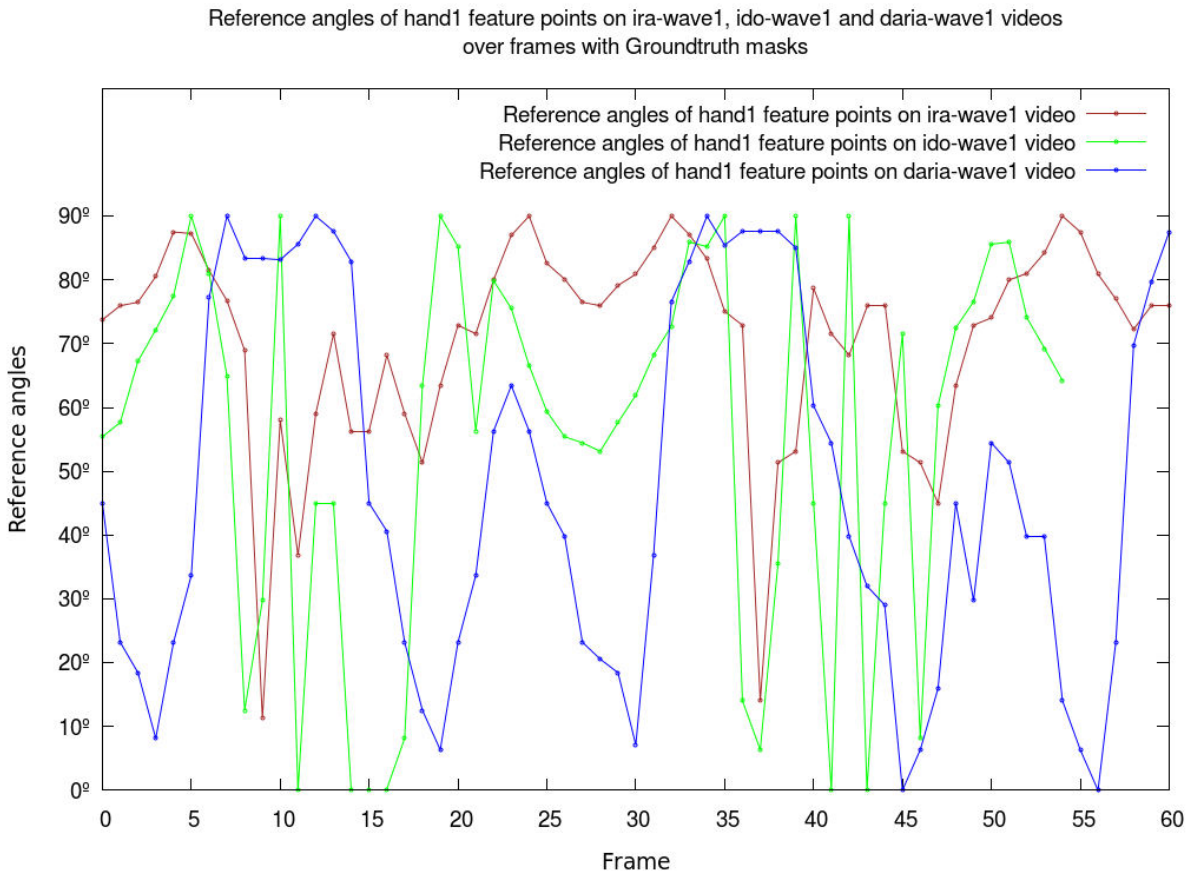


Figure 4.32: Left hand reference angles for "wave1" action videos of subjects "ido", "daria" and "ira" using Ground-truth masks

The proposed angles may indeed present themselves as interesting inputs for a classification system, since in certain actions they do provide similar patterns. On the other hand, the feature point detection can still be improved, specially when it comes to the hands, in order to also improve the angle detection. The more accurate the feature point detection is, the more precise the angle pattern for each movement would be.

4.7 - Discussion

Results show that the proposed algorithm is able to correctly identifying three out of the five anatomic points proposed by [44]. In particular, the head and feet are correctly identified on more than 90% of the cases, while the hands achieve a lower performance with a precision of approximately 45%.

The method proposed by Aggarwal [44] did not permitted the identification of the feature points. Being that one of the contributions of this work, the knowledge of the locations of the five feature points can be valuable for a classification system in addition with the proposed four angles.

Even though the algorithm performed well for feet and head feature points, the hands detection rates are not satisfactory. There are many reasons for this fact, which are addressed mainly in Section 4.2. Resolving those issues merely based on the silhouette contours is a hard challenge. Mostly because of the occlusion of the hands on the inside of the contours. One way of getting

around this issue would be collect other pixel information like contrast and also take in consideration the pixels inside the contours. Alternatively, the feature points could be tracked using a simple motion model based on a Kalman filter for example.

Additionally, anatomic measures could be used to perceive detection errors. For instance, assume a maximum length for the arm which would be calculated depending on the height of the subject. Then it would be postulated that the distance between the hands feature points and the beginning of the arms, could not be higher than the defined maximum length of the arm. This approach could identify bad hand detections where the error would be considerable. This proposal is illustrated in Figure 4.33.

The problem would be to estimate the location of the beginning of the arm which would represent the center of the circumference that delimitates the reach of each of the hands. It could be estimated by considering a point belonging to the straight line connecting the head and hand feature point, which is represented in Figure 4.33. On the other hand, temporal information could be addressed as well *i.e.* the detected feature points of past frames. If the locations in nearby frames of the same feature point are quite set apart, then there is a greater chance of error.

From another perspective, one can perceive the detection-only version of the algorithm results presented in Section 4.3 as the capability of improvement without computing any more data other than the already obtained five feature points. Improved results using this evaluation method mean that in certain cases, there is at least one feature point that if reassigned to another, could produce a lower distance error.

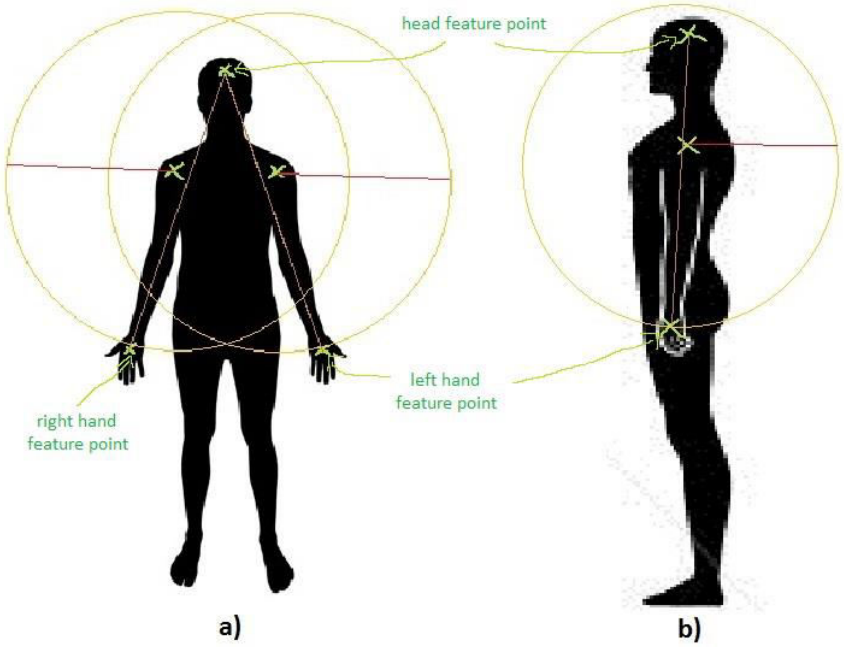


Figure 4.33: Limitation of hands location a) Front silhouette; b) Side silhouette

One major restraint that was verified in the experiments was the defective foreground segmentation arising from the background subtraction. Many individual cases where it was verified

are addressed in Section 4.5. Nonetheless, the factors that directly influence the performance of the chosen background subtraction method are majorly:

- The contrast between the desired foreground objects and the background;
- The quality of the video;
- The motion of the foreground objects relatively to the background.

The quality of the video may indeed help to provide a more refined contrast between the foreground and background. For instance, in the majority of the videos, the subjects' hands were not included in the foreground, because they were too light to stand out from it. Consequently, even if a fairly close location would be computed the detected feature point, there was always an additional error that would be avoided if the hands were correctly associated with the foreground. On the other hand, it can be seen as a trade-off *i.e.* it would have to be evaluated how much better the results would be by simply apply the algorithm to better quality videos. Nevertheless, background subtraction was not the focus of this work.

Finally, in Section 4.6 the proposed angles are evaluated both from an error measuring perspective and from an anatomic measure perspective. They can serve as a bridge to future work regarding classification of actions. On the other hand, dependently on the action, there may be feature points that present a more characteristic angle pattern over the frames *i.e.* on waving actions, the angles of the feet would be near 0º and the waving hands' angles would increase as the hand would rise until the head and decrease on the downward movement, hence providing a more distinctive variation.

4.8 - Summary

In this chapter, the conducted experiments' conditions and the corresponding results were presented. The algorithm proposed in Chapter 4 was tested in different scenarios and the results were analyzed in different perspectives. Limitations were identified and their cause isolated *i.e.* there were errors whose main cause was the defective background subtraction and not the algorithm itself. The impact of the background subtraction was found to be of importance in some cases, specially in when the colors of the subject's clothes were more similar to the background's colors. The detection-only version of the algorithm shows that there is margin for improvement of the matching criteria since it shows the best possible matching results with the detected feature points. Indeed, cases where a bad matching was made could be corrected by re-assigning the detected feature points. On the other hand, the proposed post-processing showed a small impact on the overall results. Moreover, the 4 proposed angles are analyzed in the context of some of the actions in Section 4.6. Patterns of these angles were recognized in the video sequences of several actions and it is proposed to use them as part of a classification system. A final discussion is presented in Section 4.7, where some key issues were addressed and additional solutions were proposed to improve the algorithm.

Chapter 5

Final Remarks

The conclusions regarding the conducted experiments aiming to test the proposed algorithm presented in Chapter 3 are discussed in Section 5.1. They are also be put into perspective towards possible applications and future investigation paths that can take advantage of this work in Section 5.2.

5.1 - Conclusions

In this thesis, it is proposed a new way of matching Human silhouette feature points to 5 major extremities (head, the two hands and feet) as well as 4 angles relating the head and hands and the feet and centroid of contours. The detection of the candidate feature points is based on the algorithm proposed by J. K. Aggarwal [44]. However, his work primarily targets the detection of Humans climbing fences and there was no need to match the feature points to certain parts of the body.

Comparably with the work of Aggarwal [44], the proposed method permits not only the detection of feature points in the Human silhouette contours, but also the identification of body extremities, namely head, hands and feet. The proposed objectives for this work included the proposal of a method capable of identifying key feature points of a subject's silhouette and the proposal of a descriptor based on those feature points. These objectives were achieved and the contributions of this thesis illustrate it:

- Proposal of a method capable of extracting feature points of a Human silhouette and matching them to body extremities;
- Proposal of an angular descriptor based on the extracted feature points that relates the hands and head and the contours centroid and feet.

Using the ground-truth videos, the matching precision for the feet and head was nearly 90% but for the hands feature points the average precision was between 40 and 50%. Several matching errors were analyzed and it was concluded that the matching was severely constrained whenever

the hands were occluded on the inside of the silhouette contours, which constantly happened in most of the videos. In actions that required the hands to perform a movement that implied their shape to be clearly defined apart from the torso, like waving the matching results were quite more satisfactory. These results suggest that in order to obtain an improved detection and matching of the hands feature points, it might be suited to combine the contours with additional pixel analysis on the inside of the silhouette.

The algorithm was tested in the perspective of asserting the impact of the foreground segmentation on the overall performance of the algorithm. When using background subtraction to obtain the foreground mask (hence, the human silhouette contours), several errors derived from it were detected. Nevertheless, the conditions of the analyzed videos were not often the ideal for an accurate background subtraction, namely the subjects did not always wore clothes with a clear distinctive color from the background, thus not providing a sufficient contrast. Additionally the resolution of the videos was low, even though it was appropriated for regular video-taping equipment. The average overall results were slightly badly influenced by a poor background subtraction, as discussed in Section 4.5. It has been concluded to be due to the uncertainty verified in several videos concerning the matching of the hands feature points, they are the feature points that present a poorer matching performance. This is explained by the fact that in most poses, the shape of the hands are actually concealed within the interior of the extracted contours of the subject. Nonetheless, in Section 4.7 an alternative method to debug an accurate hand detection is proposed, namely considering the possible motion area of the arm being the beginning of the arms approximated by the feature point of the head and the length of the arm calculated as a factor of the subject's height. Even though the background subtraction was indeed a restraint to an effective application of the algorithm, it does not seem to be the cause of particularly poor results in the hands.

Moreover, the algorithm was tested in order to obtain the best possible matching results with the detected points *i.e.* consider for evaluation the nearest detected point to the reference location of each of the 5 feature points, as addressed in Section 4.3. This evaluation was performed in order to assert the improvement margin of the algorithm by simply re-assigning the detected feature points in order to obtain a better match. The results showed some improvement, though not very high, which suggests that there is some room to work on the final 5 feature points re-matching.

Additionally, the proposed post-processing was concluded to have little impact on the overall results. However, it collects a different set of feature points that has the potential to indeed improve the matching results. Namely, the feature points of the Freeman chain code method in the neighborhood of each of the five feature points of the proposed algorithm could be collected and analysed in order to evaluate the anatomic structure around them. Nevertheless, the criteria that was chosen in this work to combine the feature points acquired from the Freeman chain code based algorithm with the outputted five feature points of the proposed algorithm revealed to be too

simplistic. Yet, a more robust criteria could be explored in order to combine both algorithms in order to achieve better results.

Based on the identification of the proposed 5 feature points, the four proposed angles relating the hands and head and feet and the centroid were evaluated. The analysis in Section 4.6 also focuses in detecting patterns in the angles verified by different subjects in the same action. Some feature points present particular angle patterns over the course of action. This suggests that they can be used to perform classification tasks, hence identifying the type of action being conducted. On the other hand, the matching performance has to be taken into consideration in order to assure an accurate measurement of the proposed angles.

Overall, the feet and head showed a consistent matching performance and despite the difficulties of detecting and matching the feature points of the hands, the reasons that led to them have been identified. In conclusion, the proposed angles seem to present themselves as a valid feature that can help identify simple actions.

5.2 - Future Work

The matching results for the feet and head were quite satisfactory. However, the hands detection did not perform as well as desired. Many reasons were determined to be the cause of these results and were discussed during Chapter 4. Henceforth, it becomes necessary to improve the algorithm in this matter. The criteria to combine the results obtained from the Freeman chain code based method proposed as a post-processing can indeed be improved. Namely, for each of the 5 detected feature points, the neighborhood feature points of the Freeman chain code based algorithm within a defined distance could be considered. After processing them in order to extract structural anatomical data, hence identifying the surrounding anatomical shape, a location of the feature point closer to its corresponding reference location could be calculated.

Moreover, even though in this thesis it was proposed a single angular descriptor, the proposed 5 feature points have the potential to provide additional descriptors. Examples include distances between hands and feet feature points, different angles, possibly directly relating the position of feet and hands. Afterwards, likewise it was discussed in Section 4.6, the capability of such descriptors to discriminate a designated action should be evaluated, since ultimately, a major future work goal would be using them in an actions classification system.

On a different perspective, a more complex experimental environment should be kept in mind in order to assess the capabilities of the algorithm in a more realistic way. Even though the experiments in which the proposed matching algorithm was tested merely contained one subject in each video, in real life situations naturally there might be several subjects. Moreover, the considered scenarios do not present complex backgrounds, not even moving objects in the background that could cause additional noise in the background subtraction.

In order to extrapolate the application of this matching algorithm to more realistic scenarios, it is proposed to combine it with a Human detection system. On the other hand, the outputs of the

proposed matching algorithm could be the inputs of a classification system of simple actions. The locations of the 5 feature points as well as the proposed 4 angles can provide valuable input in order to identify simple actions.

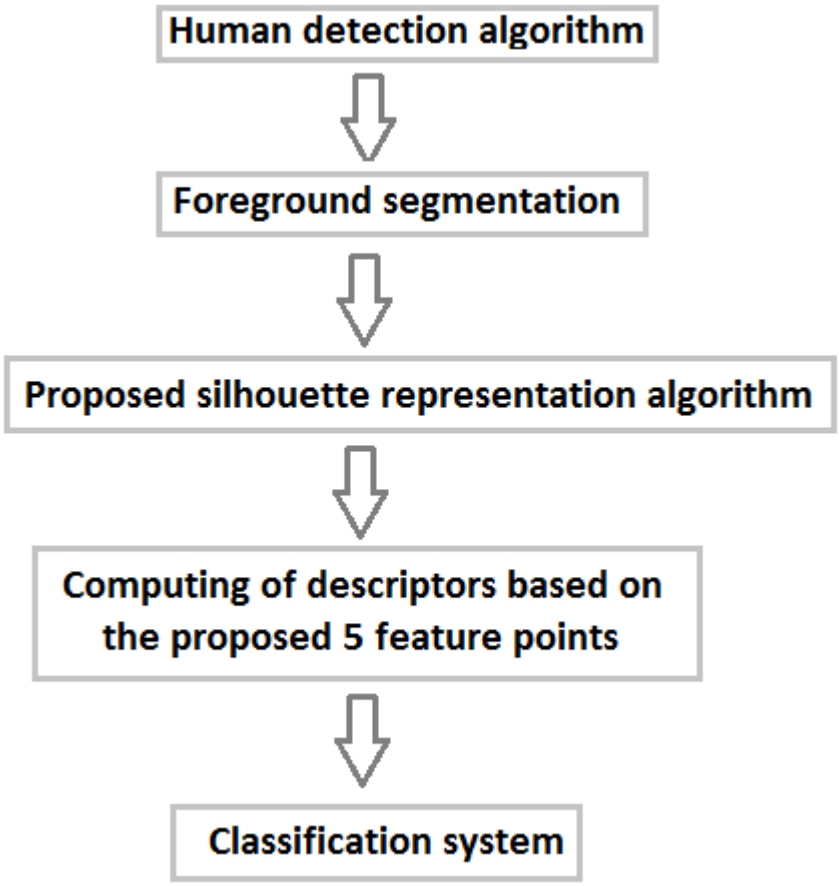


Figure 5.1: Proposed integration of the matching algorithm

The Human detection algorithm would output the section of the video corresponding to each Human individual separately, which then would have to be submitted to a foreground segmentation method like background subtraction. Obviously this implies the tracking of each individual, which would be an extra task of the Human detection algorithm. Subsequently, the proposed matching algorithm would process the silhouette contours and provide the necessary input features for the classification.

On the other hand, the actual method used to compute the five feature points could be re-designed considering a Skeletonization approach which is addressed in Section 2.3. Instead of considering the highest point of contours and the centroid to calculate the distances to each contour points, other points could be considered and other paradigms could be deliberated to estimate a more accurate location of the five feature points.

References

[1] R. Graves, *The Greek Myths: Illustrated Edition*. Penguin Books, 1992, p. 224.

[2] M. D. Bowan, "INTEGRATING VISION WITH THE OTHER SENSES," 2012.

[3] F. Cities, "Future Cities: An Ecosystem for the Future," 2013. [Online]. Available: http://www.futurecities.up.pt/site/.

[4] A. Andreopoulos, S. Hasler, H. Wersing, H. Janssen, J. K. Tsotsos, and E. Korner, "Active 3D Object Localization Using a Humanoid Robot," *IEEE Trans. Robot.*, vol. 27, no. 1, pp. 47–64, Feb. 2011.

[5] C. Wei and M. Anlong, "Hierarchical filling based object segmentation for silhouette extraction from surveillance videos," in *2009 2nd IEEE International Conference on Broadband Network & Multimedia Technology*, 2009, pp. 580–585.

[6] I. Zafar, U. Zakir, I. Romanenko, R. M. Jiang, and E. Edirisinghe, "Human silhouette extraction on FPGAs for infrared night vision military surveillance," in *2010 Second Pacific-Asia Conference on Circuits, Communications and System*, 2010, vol. 1, pp. 63–66.

[7] Z. Liu, K. T. Ng, S. C. Chan, and X.-W. Song, "A new multi-view articulated human motion tracking algorithm with improved silhouette extraction and view adaptive fusion," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, 2013, pp. 713–716.

[8] L. Changzheng, L. Zhiguang, H. Zhiheng, and Y. Guiyun, "Depth and silhouette based skeletal joints tracking," *Int. J. Adv. Comput. Technol.*, vol. 3, no. 11, pp. 72–79, 2011.

[9] M. Milanova, L. Bocchi, and T. Stoev, "Video-based human motion estimation system for gaming user interface," in *2009 International IEEE Consumer Electronics Society's Games Innovations Conference*, 2009, pp. 37–42.

[10] Y. Koyama, T. Matsumoto, Y. Shirai, N. Shimada, and M. Ueda, "Silhouette Extraction based on Time Sequence Histograms and Graph Cut for Golf Swing Diagnosis," *IEEJ Trans. Electron. Inf. Syst.*, vol. 132, no. 11, pp. 1840–1846, 2012.

[11] P. Foucher, D. Moreno Eddowes, and J. Lopez Krahe, "Traffic light silhouettes recognition using fourier descriptors," in *Proceedings of the 5th IASTED International Conference on Visualization, Imaging, and Image Processing, VIIP 2005*, 2005, pp. 186–190.

[12] Y. Shen, W. Hu, J. Liu, M. Yang, B. Wei, and C. T. Chou, "Efficient background subtraction for real-time tracking in embedded camera networks," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems - SenSys '12*, 2012, p. 295.

[13] Y. Benezeth, B. Emile, H. Laurent, and C. Rosenberger, "Comparative Study of Background Subtraction Algorithms," vol. 19, pp. 1–30, 2010.

[14] Z. Wang, H. Xu, L. Sun, and S. Yang, “Background subtraction in dynamic scenes with adaptive spatial fusing,” *2009 IEEE Int. Work. Multimed. Signal Process.*, pp. 1–6, Oct. 2009.

[15] A. Amato, M. Mozerov, I. Huerta, J. Gonzàlez, and J. J. Villanueva, “Background Subtraction Technique based on Chromaticity and Intensity Patterns,” pp. 6–9, 2008.

[16] G. Mori, S. Belongie, and J. Malik, “Efficient shape matching using shape contexts.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1832–7, Nov. 2005.

[17] Linda G. Shapiro and George C. Stockman, *Computer Vision [Paperback]*. Prentice Hall; 1 edition, 2001, p. 608.

[18] R. Achanta, S. Hemami, F. Estrada, S. Sabine, and D. L. Epfl, “Frequency-tuned Salient Region Detection,” no. Ic, pp. 1597–1604, 2009.

[19] A. Mondal, S. Ghosh, and A. Ghosh, “Efficient silhouette based contour tracking,” *2013 Int. Conf. Adv. Comput. Commun. Informatics*, pp. 1781–1786, Aug. 2013.

[20] Z. Lin and L. S. Davis, “Shape-based human detection and segmentation via hierarchical part-template matching.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 604–18, Apr. 2010.

[21] Y. Tan, Y. Guo, and C. Gao, “Background subtraction based level sets for human segmentation in thermal infrared surveillance systems,” *Infrared Phys. Technol.*, vol. 61, pp. 230–240, Nov. 2013.

[22] T. Lindeberg, *Encyclopedia of Mathematics*. Springer, 2001.

[23] A. L. IÂrbus and A. L. IÂrbus, *Eye movements and vision*. 1967.

[24] G. Titelman, *Random House Dictionary of America’s Popular Proverbs and Sayings: Second Edition*. Random House Reference, 2000, p. 496.

[25] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[26] M. K. E. R. Simone Frintrop, “A real-time visual attention system using integral images.”

[27] Y.-F. Ma and H.-J. Zhang, “Contrast-based image attention analysis by using fuzzy growing,” in *Proceedings of the eleventh ACM international conference on Multimedia - MULTIMEDIA ’03*, 2003, p. 374.

[28] Y. Hu, X. Xie, W.-Y. Ma, L.-T. Chia, and D. Rajan, *Advances in Multimedia Information Processing - PCM 2004*, vol. 3332. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 993–1000.

[29] X. Hou and L. Zhang, “Saliency Detection: A Spectral Residual Approach,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[30] C. K. P. P. Jonathan Harel, “Graph-based visual saliency.”

[31] R. Poppe and M. Poel, “Example-based pose estimation in monocular images using compact fourier descriptors.” University of Twente, Centre for Telematics and Information Technology, 01-Mar-2005.

[32] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[33] Ming-Kuei Hu, “Visual pattern recognition by moment invariants,” *IEEE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.

[34] C.-C. Chen, “Improved moment invariants for shape discrimination,” *Pattern Recognit.*, vol. 26, no. 5, pp. 683–686, May 1993.

[35] J. M. Joo, “Boundary geometric moments and its application to automatic quality control in the Industry,” 2006.

[36] C. Chen, Y. Zhuang, and J. Xiao, “Silhouette representation and matching for 3D pose discrimination – A comparative study,” *Image Vis. Comput.*, vol. 28, no. 4, pp. 654–667, Apr. 2010.

[37] a. Elgammal, “Inferring 3D body pose from silhouettes using activity manifold learning,” *Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, 2004. CVPR 2004.*, vol. 2, pp. 681–688, 2004.

[38] W. Niblack and J. Yin, “A pseudo-distance measure for 2D shapes based on turning angle,” in *Proceedings., International Conference on Image Processing*, 1995, vol. 3, pp. 352–355.

[39] E. M. Arkin, L. P. Chew, D. P. Huttenlocher, and K. Kedem, “34; I-22 Ma 18.”

[40] N. R. Howe, “Silhouette Lookup for Monocular 3D Pose Tracking,” no. February 2006.

[41] L. Gorelick, M. Galun, E. Sharon, and R. Basri, “Shape Representation and Classi fi cation Using the Poisson Equation.”

[42] P. a. Tresadern and I. Reid, “An Evaluation of Shape Descriptors for Image Retrieval in Human Pose Estimation,” *Procedings Br. Mach. Vis. Conf. 2007*, pp. 65.1–65.10, 2007.

[43] H. Fujiyoshi and A. J. Lipton, “Real-time human motion analysis by image skeletonization,” in *Proceedings Fourth IEEE Workshop on Applications of Computer Vision. WACV’98 (Cat. No.98EX201)*, 1998, pp. 15–21.

[44] E. Yu, J. K. Aggarwal, and A. Tx, “Detection of Fence Climbing from Monocular Video,” pp. 18–21, 2006.

[45] H.-S. Chen, H.-T. Chen, Y.-W. Chen, and S.-Y. Lee, “Human action recognition using star skeleton,” in *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks - VSSN ’06*, 2006, p. 171.

[46] E. Yu and J. K. Aggarwal, “Human action recognition with extremities as semantic posture representation,” *2009 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 1–8, Jun. 2009.

[47] J. J. V. W. Ru Telea, “An augmented fast marching method for computing skeletons and centerlines.”

[48] R. L. Ogniewicz and O. Kübler, “Hierarchic Voronoi skeletons,” *Pattern Recognit.*, vol. 28, no. 3, pp. 343–359, Mar. 1995.

[49] H. Blum and R. N. Nagel, “Shape description using weighted symmetric axis features,” *Pattern Recognit.*, vol. 10, no. 3, pp. 167–180, Jan. 1978.

[50] L. Jiang, J. Yao, B. Li, F. Fang, Q. Zhang, and M. Q. Meng, “Automatic Body Feature Extraction from Front and Side Images,” vol. 2012, no. December, pp. 94–100, 2012.

[51] S. Avidan, “Fast Human Detection Using a Cascade of Histograms of Oriented Gradients,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR’06)*, 2006, vol. 2, pp. 1491–1498.

[52] O. Tuzel, F. Porikli, and P. Meer, “Human Detection via Classification on Riemannian Manifolds,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[53] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, “Human detection using partial least squares analysis,” in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 24–31.

[54] H. Freeman, “On the Encoding of Arbitrary Geometric Configurations,” *IEEE Trans. Electron. Comput.*, vol. EC-10, no. 2, pp. 260–268, Jun. 1961.

[55] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as Space-Time Shapes,” in *The Tenth IEEE International Conference on Computer Vision (ICCV’05)*, 2005, pp. 1395–1402.

Appendix A

Additional results

The conducted experiments originated several plots that permitted a visual representation of the obtained results. Some of them were used during Chapter 4, however most of them were not needed to illustrate the discussion and analysis that took place. In this Appendix, a more comprehensive sample of the generated plots are accessible. Particularly, the average precision, point-to-point and to nearest point Euclidean distance and the angle error for each feature point individually. Three main experimental scenarios are considered:

- Background subtraction vs Ground-truth videos for foreground segmentation;
- Usage of the post-processing vs no post-processing;
- Euclidean distance to reference point vs Euclidean distance to nearest point.

Nonetheless, it is important to keep in mind that these are average results divided by category of action and feature point. For each scenario, and for each of the 30 processed videos much more data was generated, which included for each individual frame the distance to reference point, the angle error, the reference angles, the precision values and every frame with a visual representation of the detected and reference points and the corresponding error.

A.1 - Background subtraction vs Ground-truth

A.1.1 - Precision

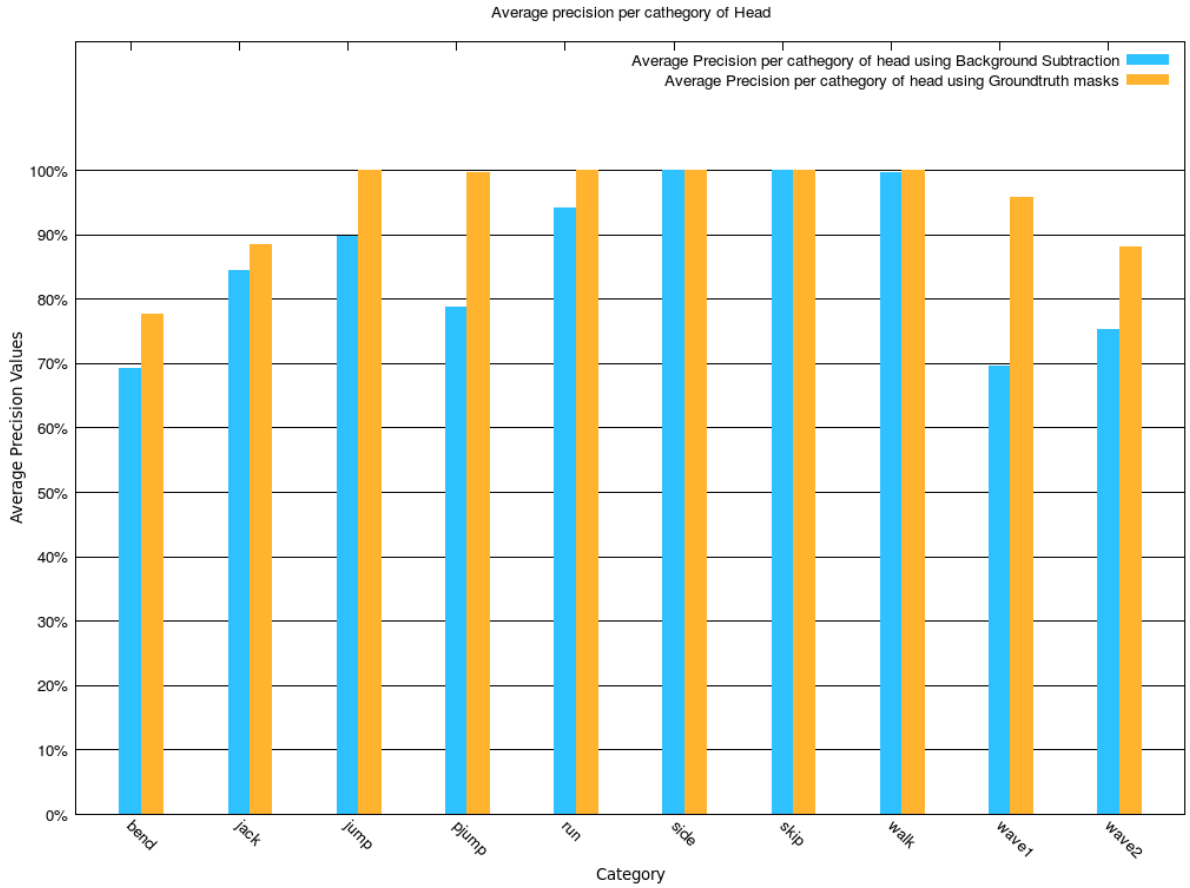


Figure A.1: Average precision for head feature point per category using Background Subtraction and Ground-truth masks

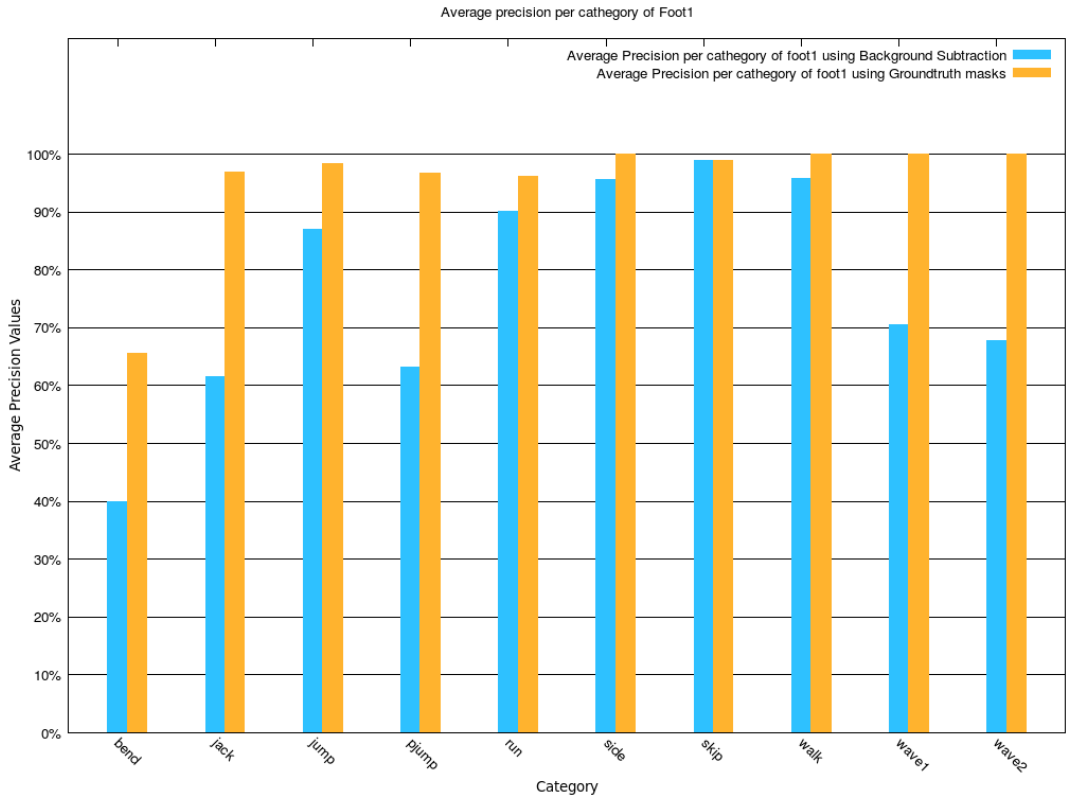


Figure A.2: Average precision for left foot feature point per category using Background Subtraction and Ground-truth masks

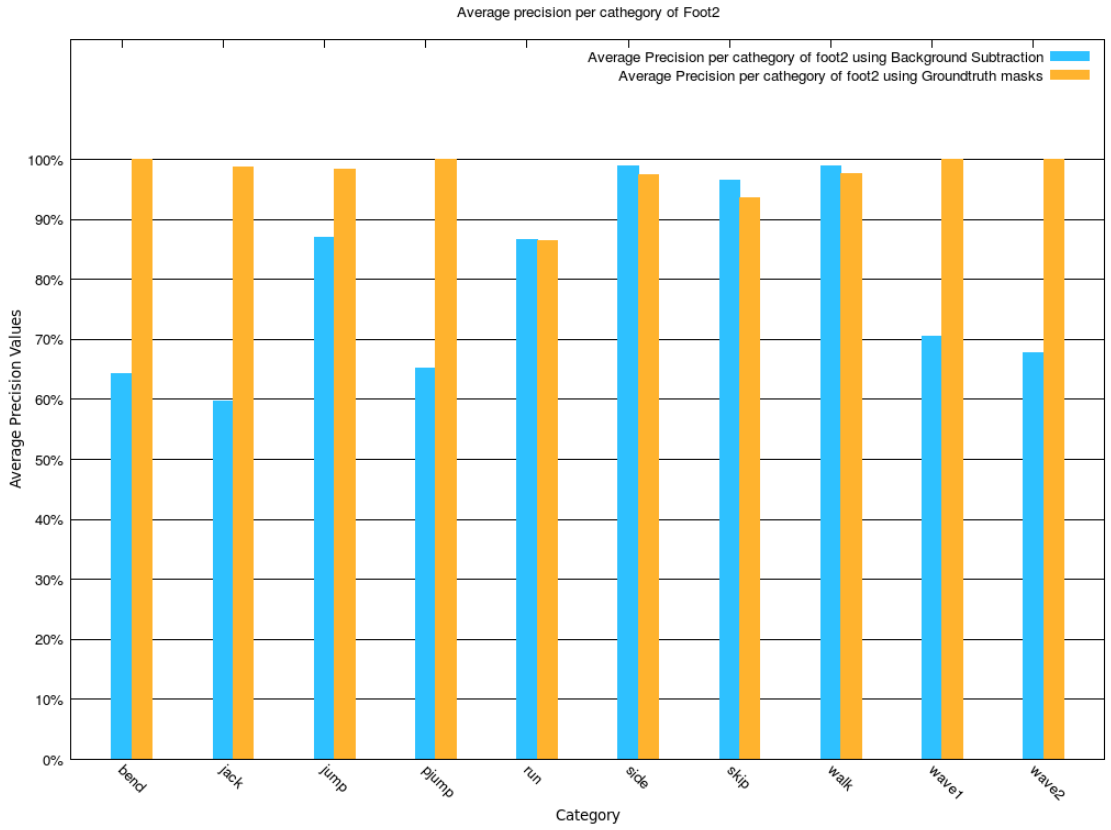


Figure A.3: Average precision for right foot feature point per category using Background Subtraction and Ground-truth masks

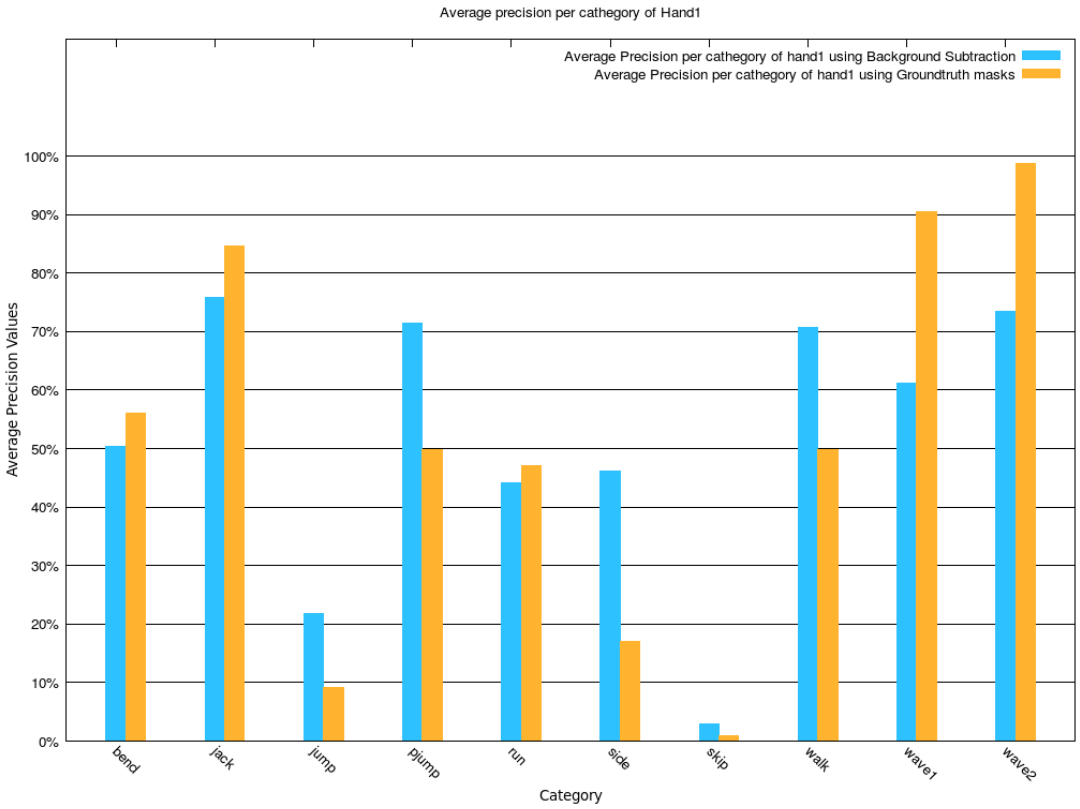


Figure A.4: Average precision for left hand feature point per category using Background Subtraction and Ground-truth masks

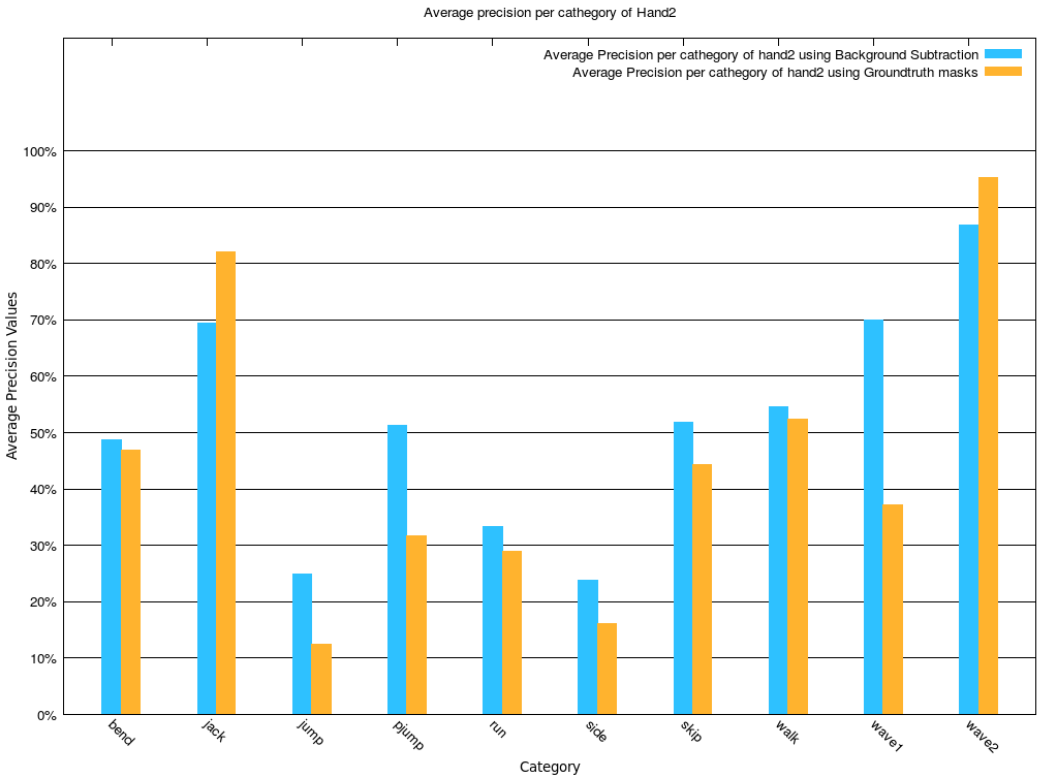


Figure A.5: Average precision for right hand feature point per category using Background Subtraction and Ground-truth masks

A.1.2 - Euclidean distance error

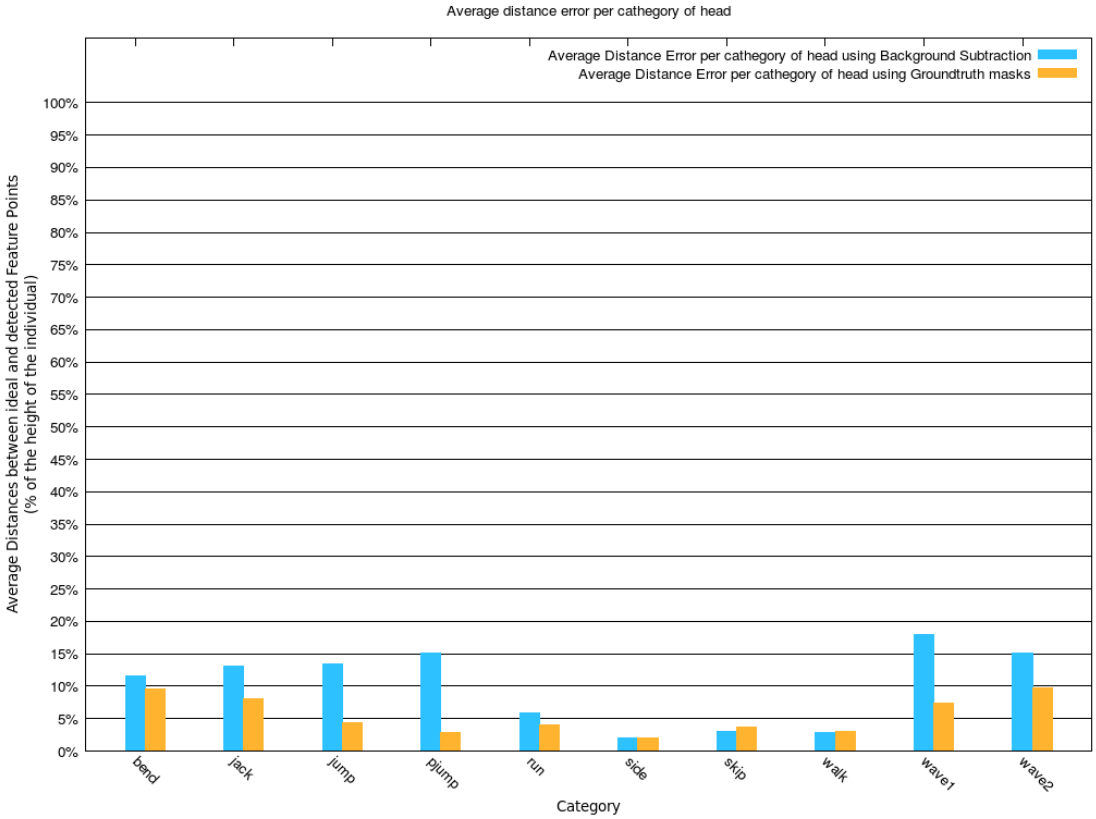


Figure A.6: Average distance to reference point for head feature point per category using Background Subtraction and Ground-truth masks

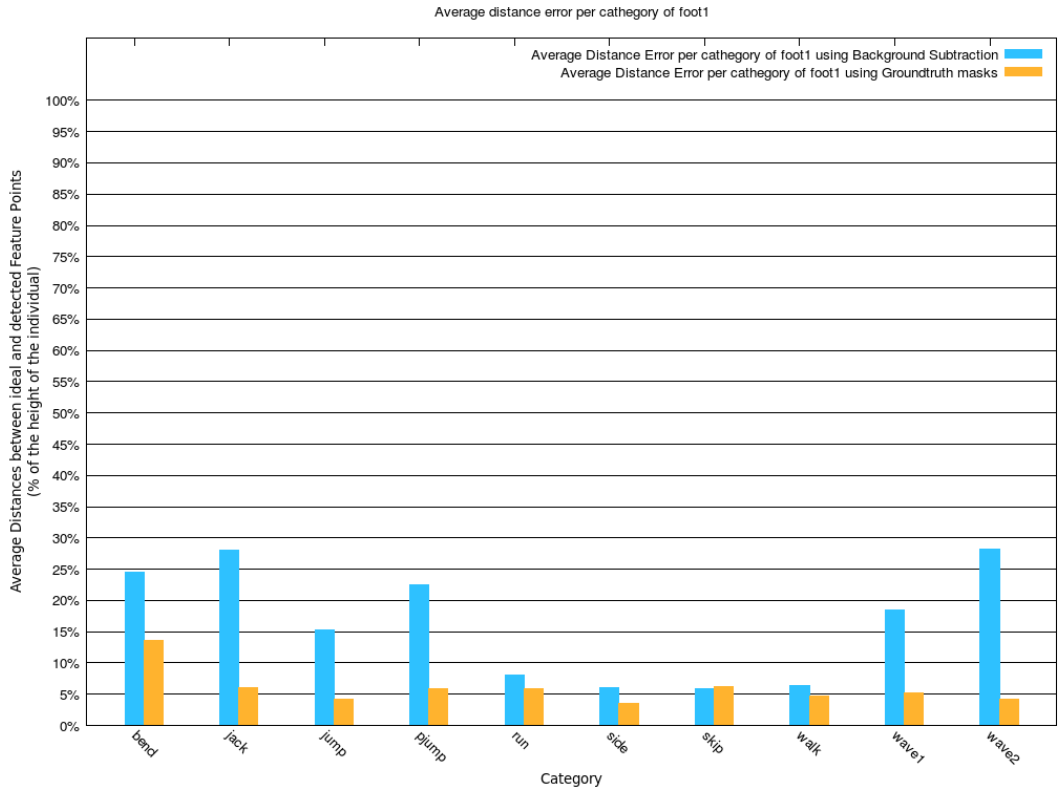


Figure A.7: Average distance to reference point for left foot feature point per category using Background Subtraction and Ground-truth masks

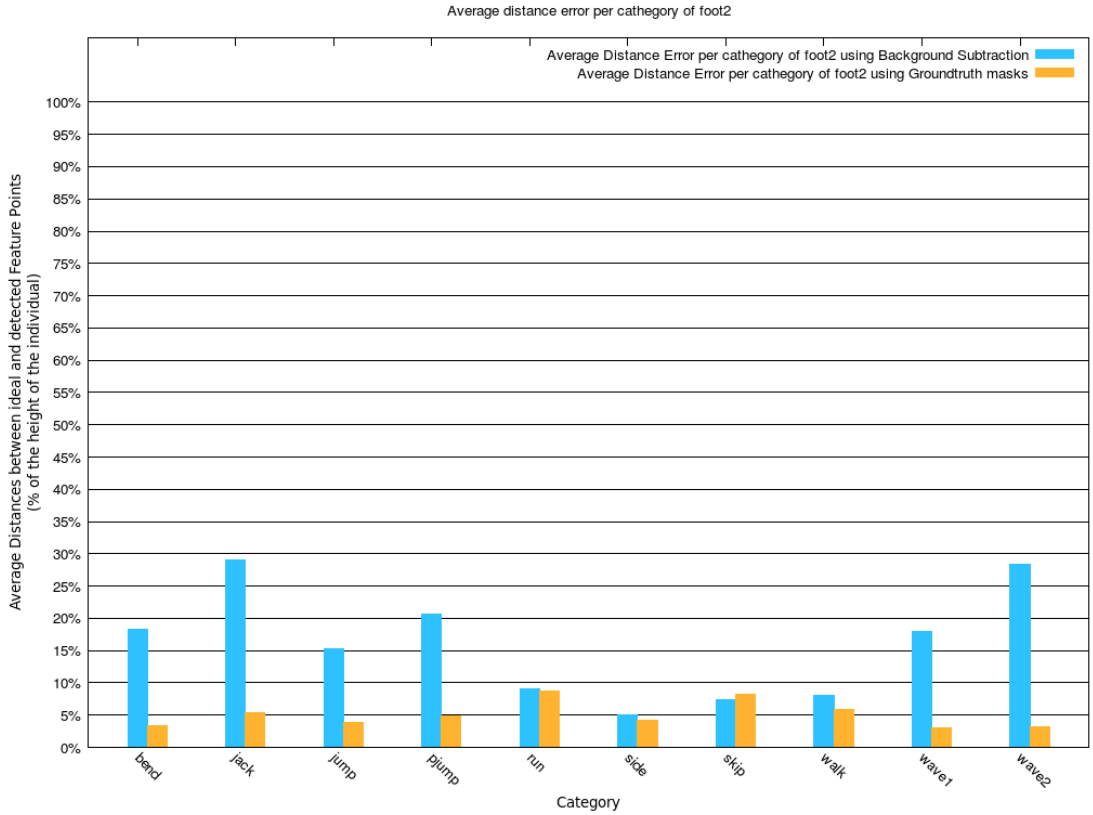


Figure A.8: Average distance to reference point for right foot feature point per category using Background Subtraction and Ground-truth masks

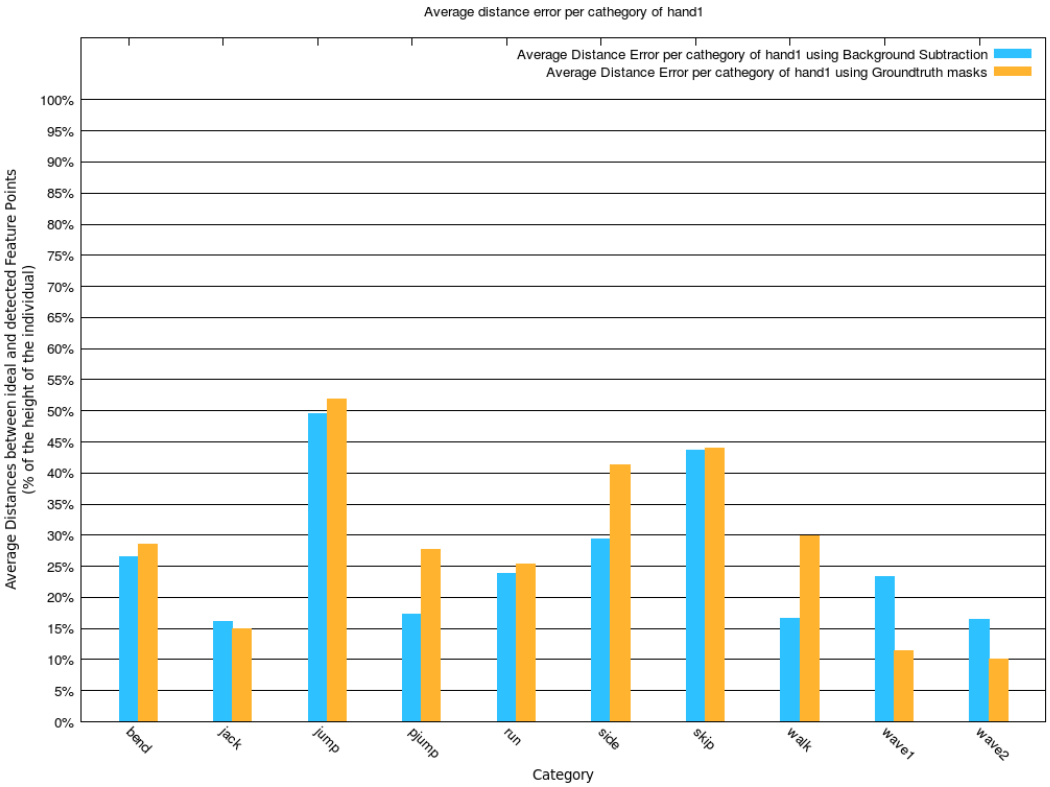


Figure A.9: Average distance to reference point for left hand feature point per category using Background Subtraction and Ground-truth masks

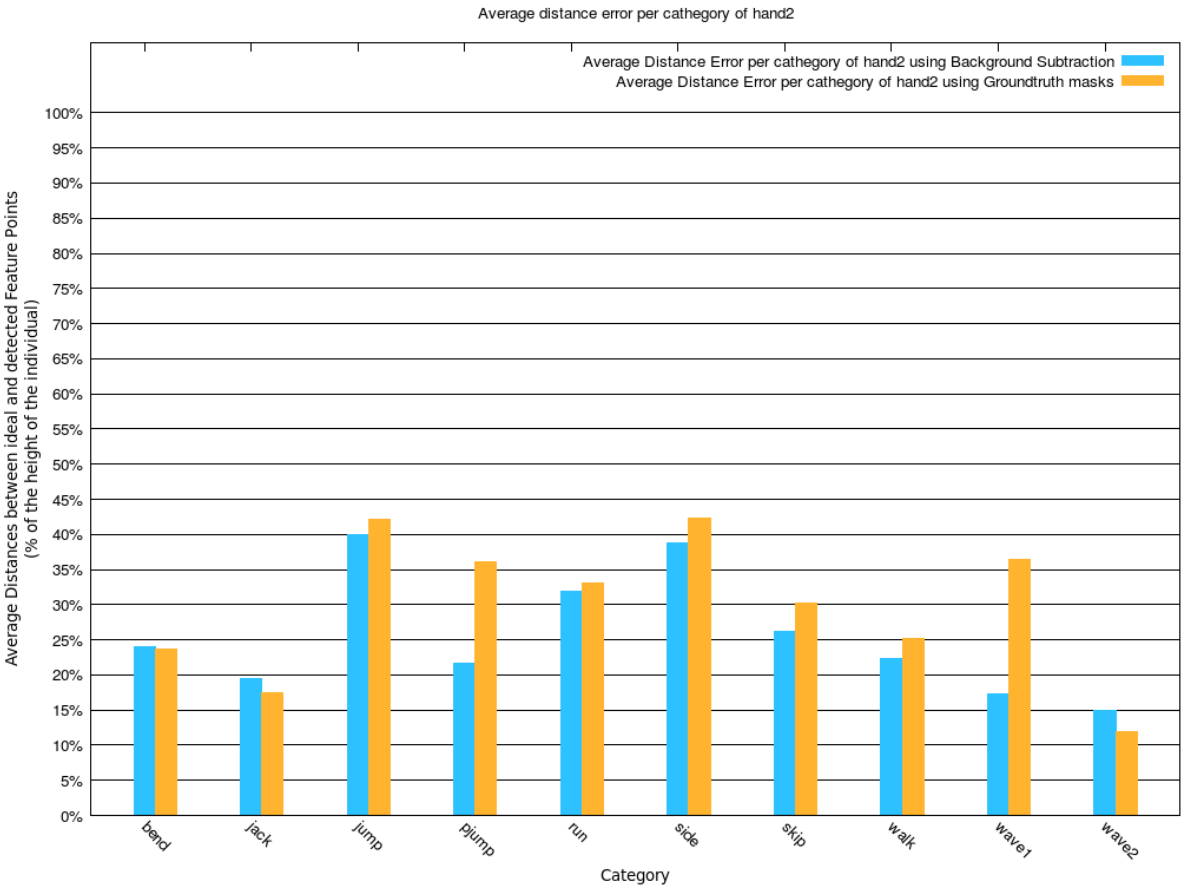


Figure A.10: Average distance to reference point for right hand feature point per category using Background Subtraction and Ground-truth masks

A.1.3 - Angle error

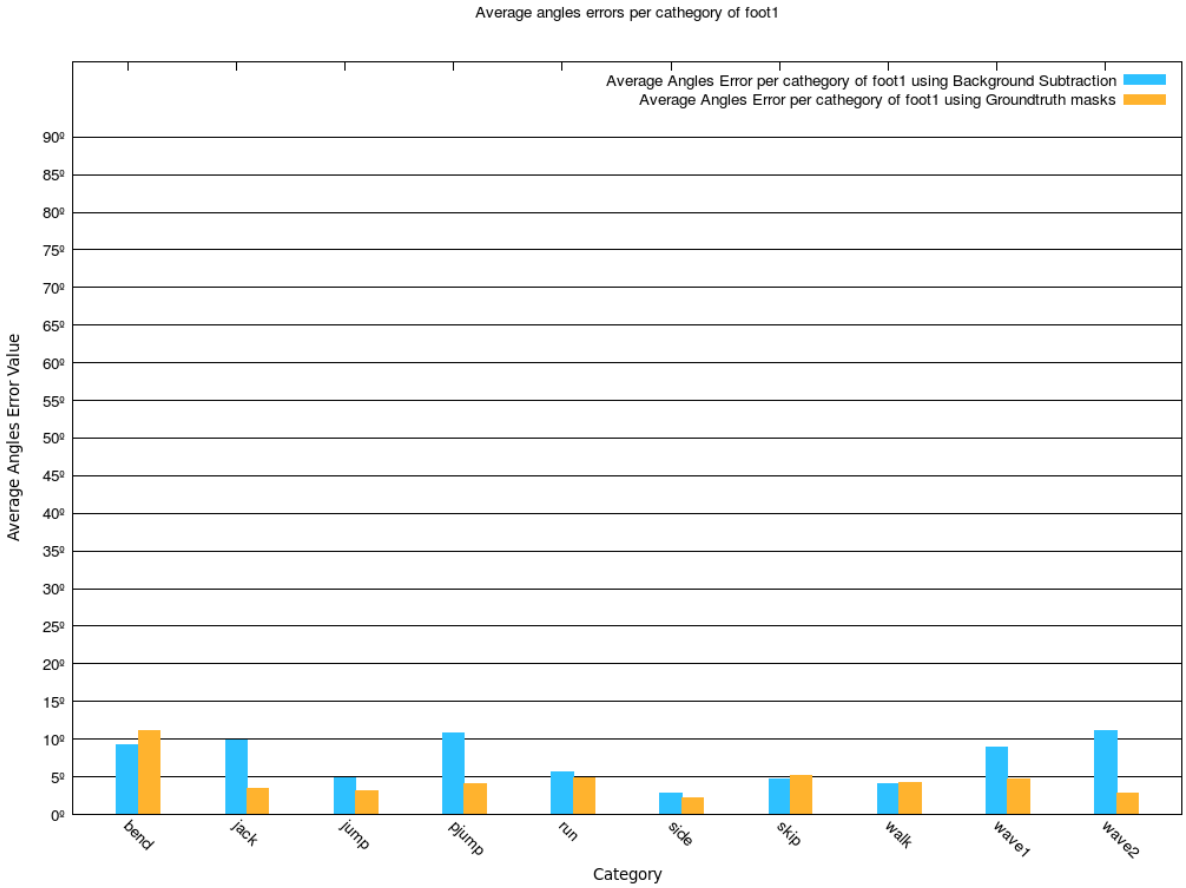


Figure A.11: Average angle error for ß1 angle per category using Background Subtraction and Ground-truth masks

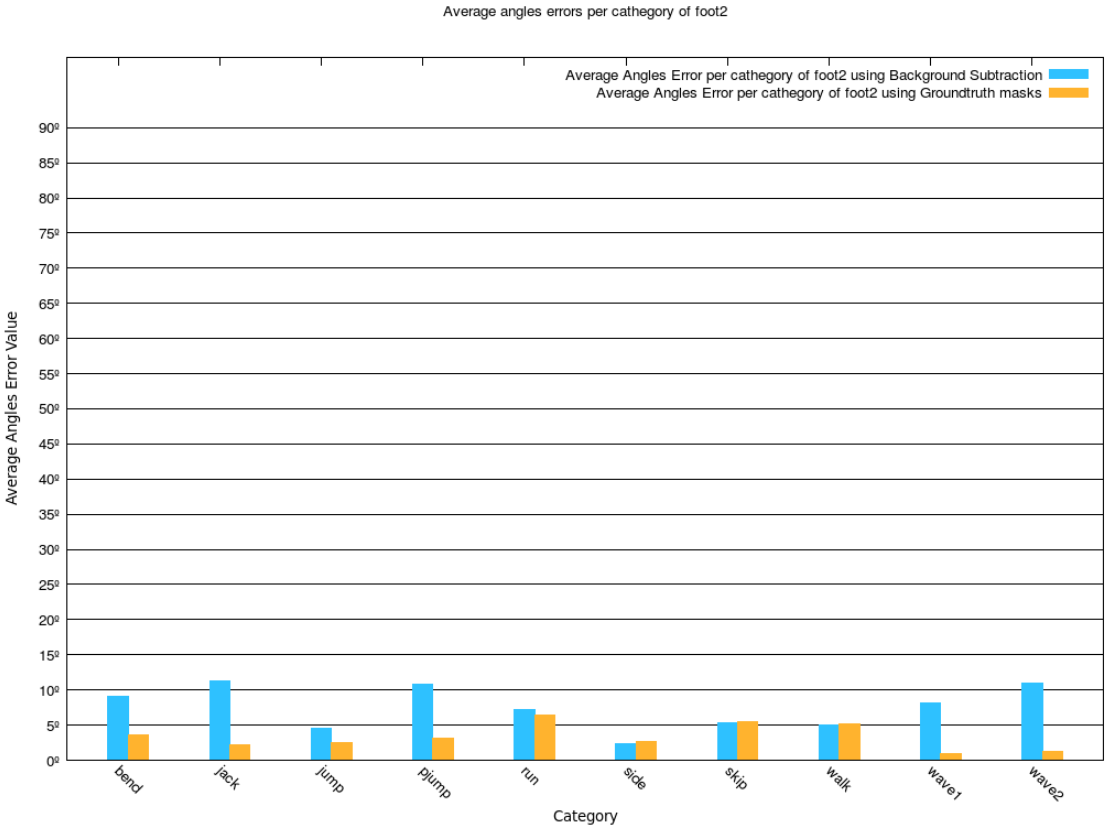


Figure A.12: Average angle error for ß2 angle per category using Background Subtraction and Ground-truth masks

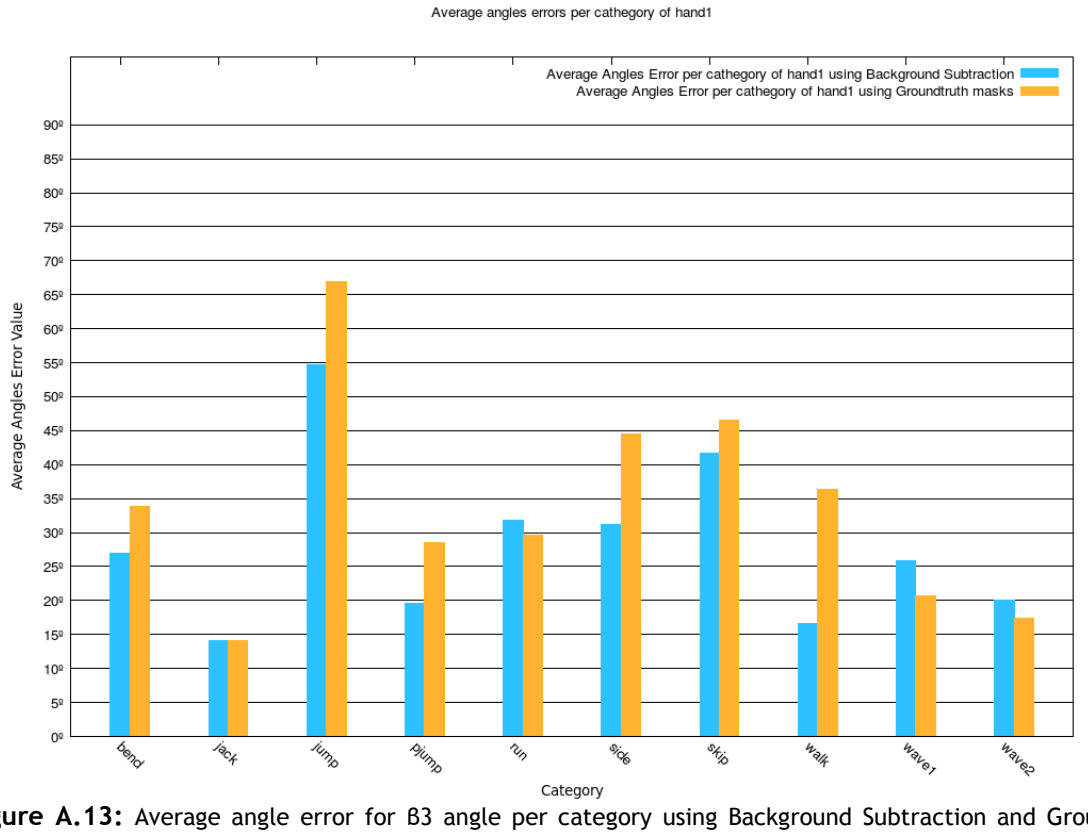


Figure A.13: Average angle error for ß3 angle per category using Background Subtraction and Ground-truth masks

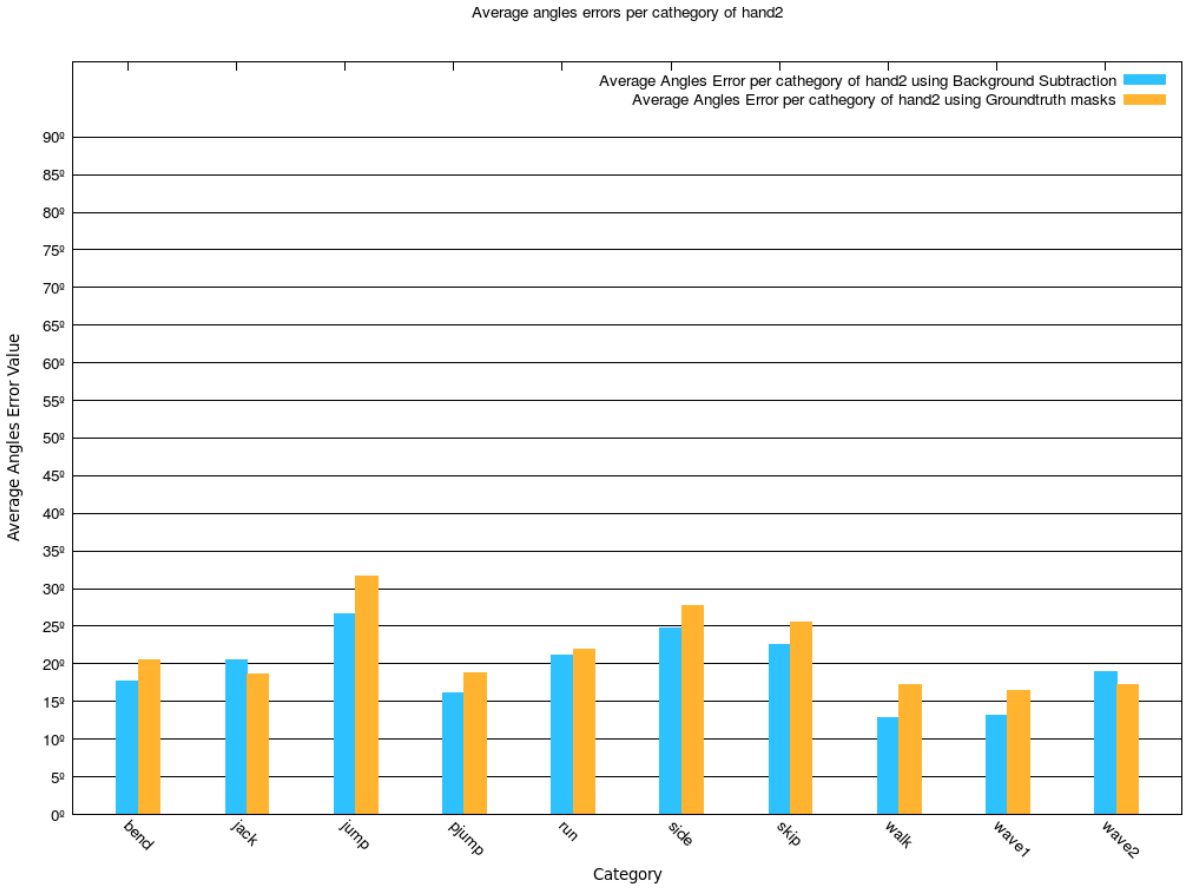


Figure A.14: Average angle error for β4 angle per category using Background Subtraction and Ground-truth masks

A.2 - Post-processing vs No post-processing

A.2.1 - Precision

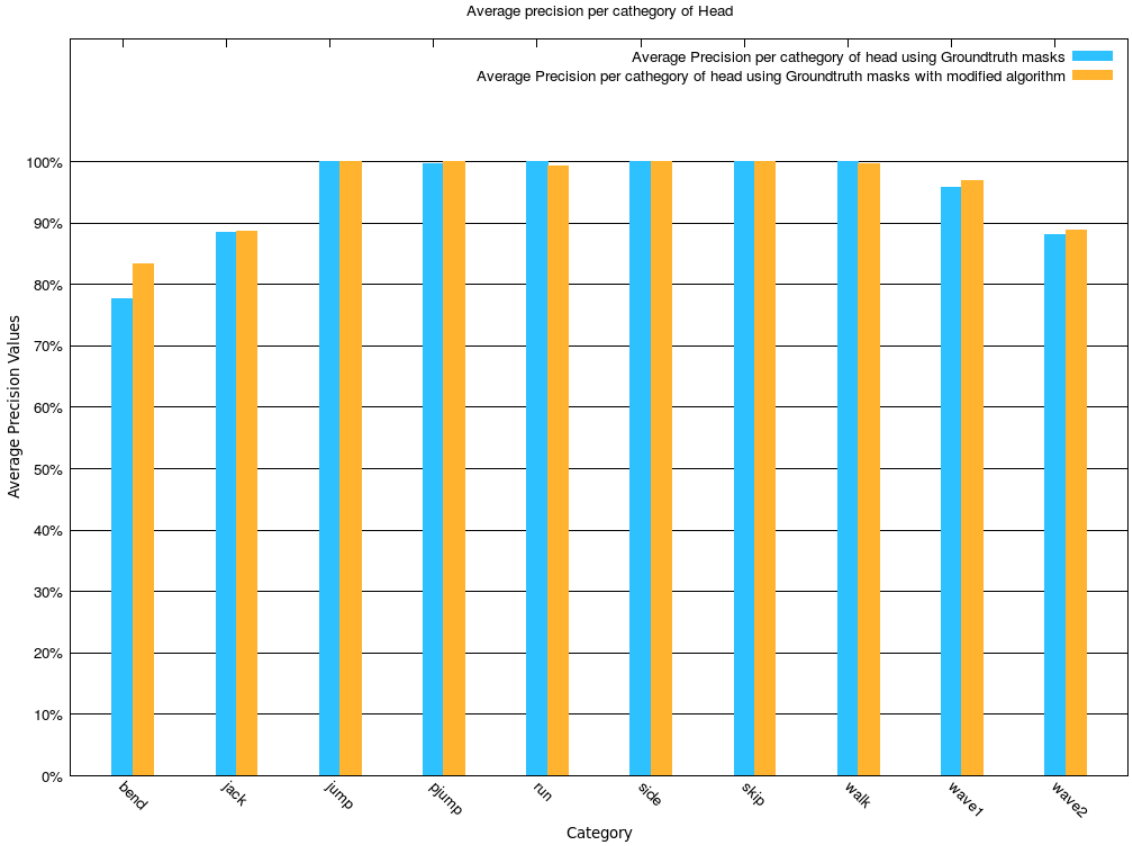


Figure A.15: Average precision for head feature point per category with and without post-processing using Ground-truth masks

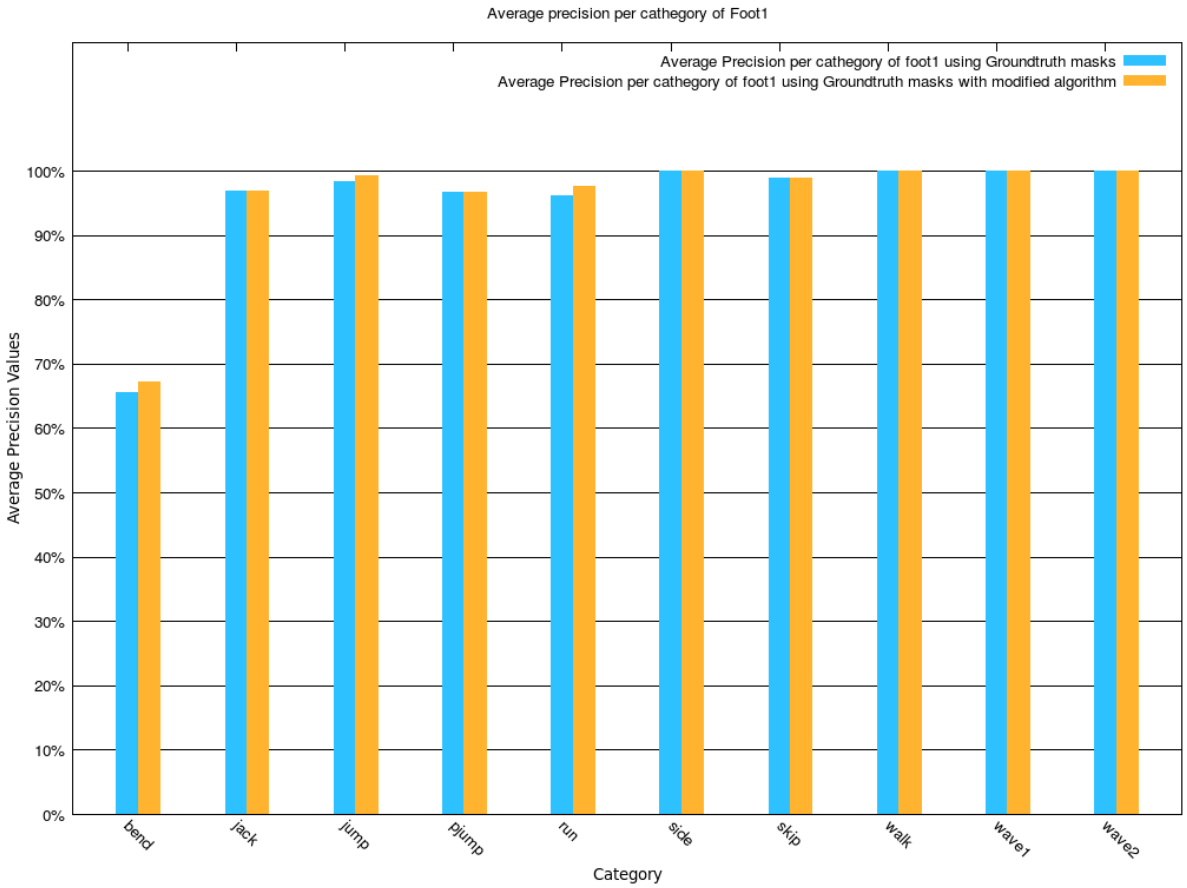


Figure A.16: Average precision for left foot feature point per category with and without post-processing using Ground-truth masks

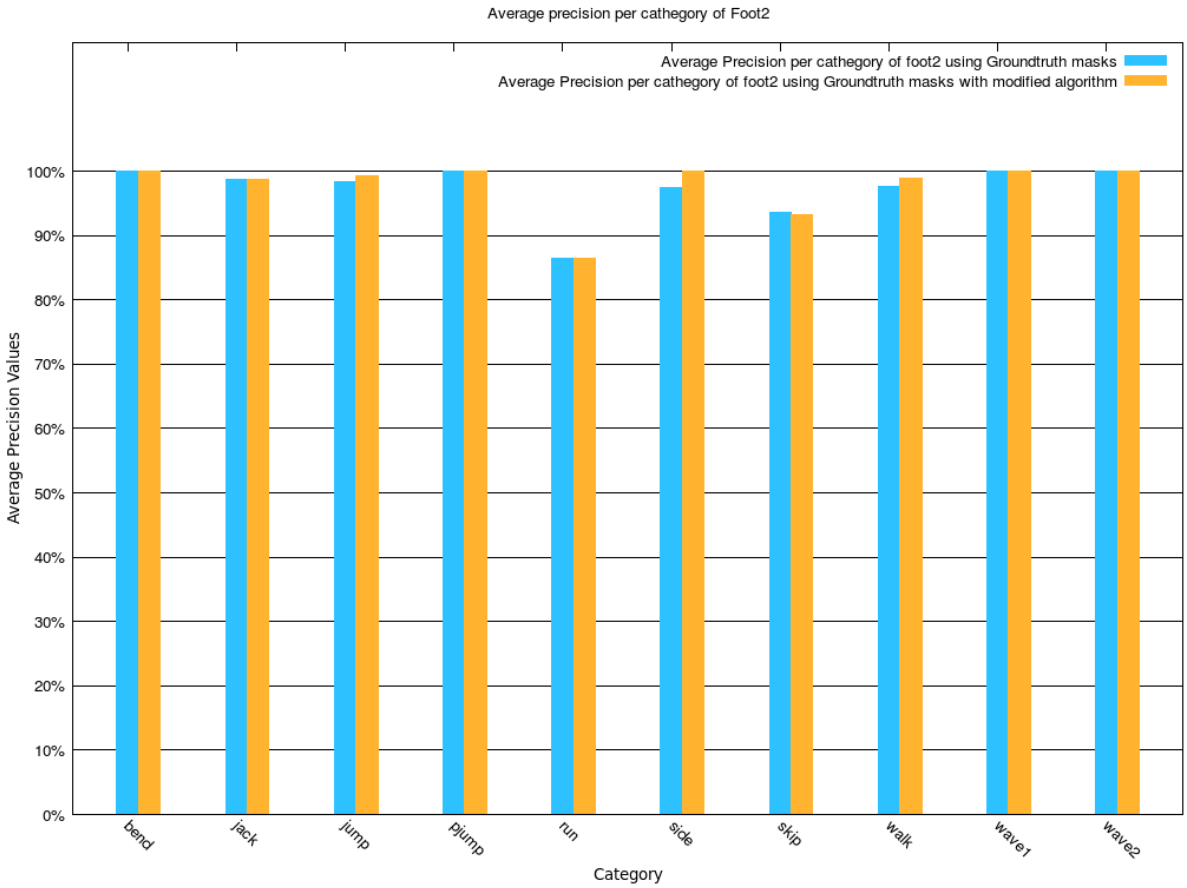
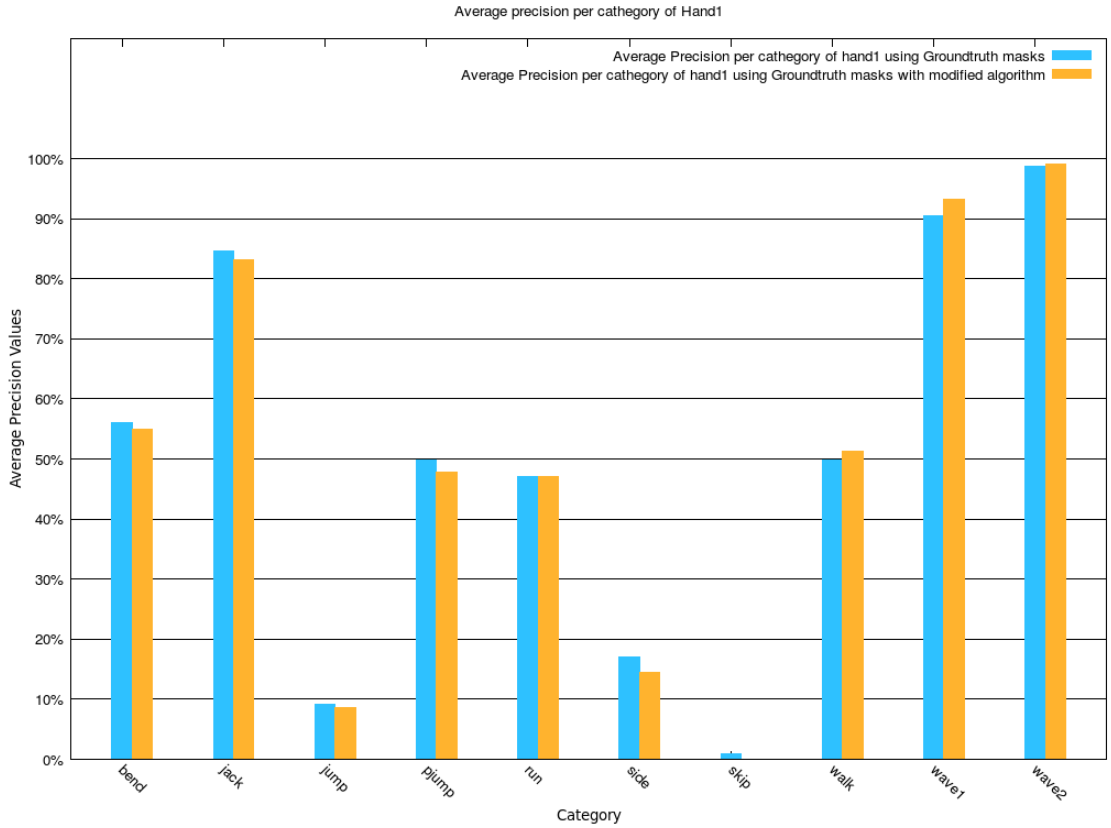


Figure A.17: Average precision for right foot feature point per category with and without post-processing using Ground-truth masks



FigureA.18: Average precision for left hand feature point per category with and without post-processing using Ground-truth masks

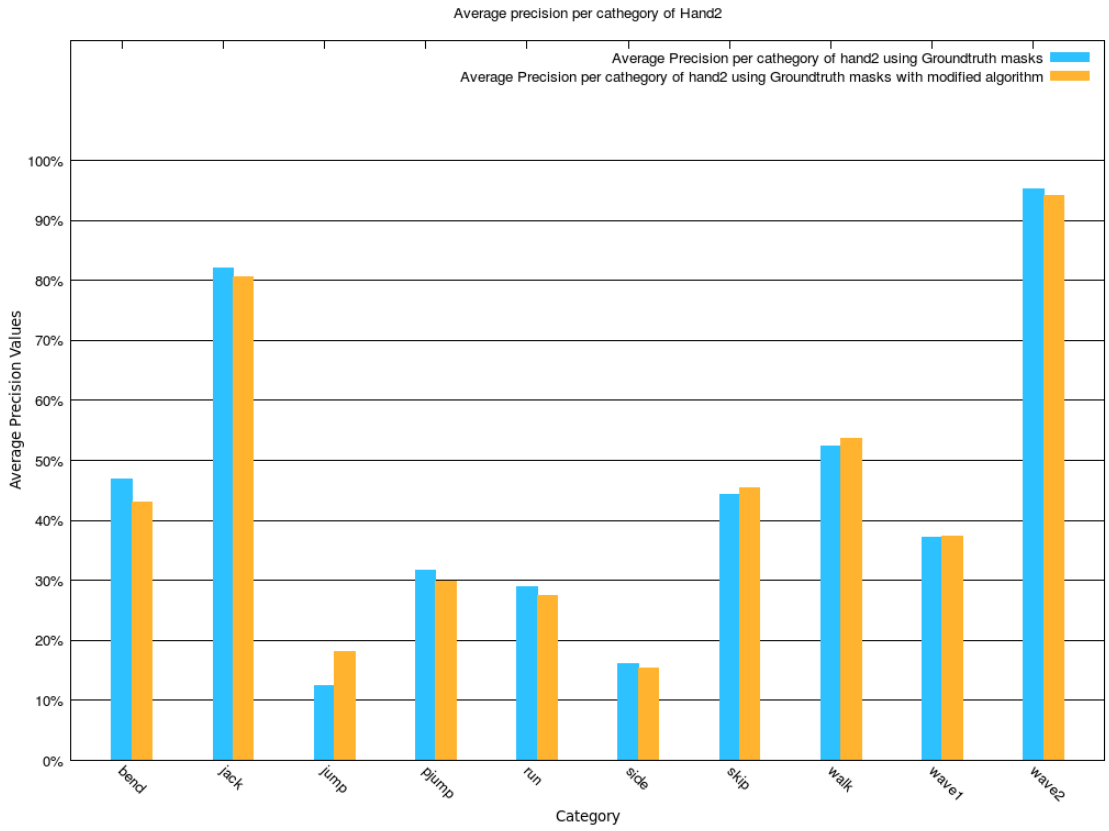


Figure A.19: Average precision for right hand feature point per category with and without post-processing using Ground-truth masks

A.2.2 - Euclidean distance error

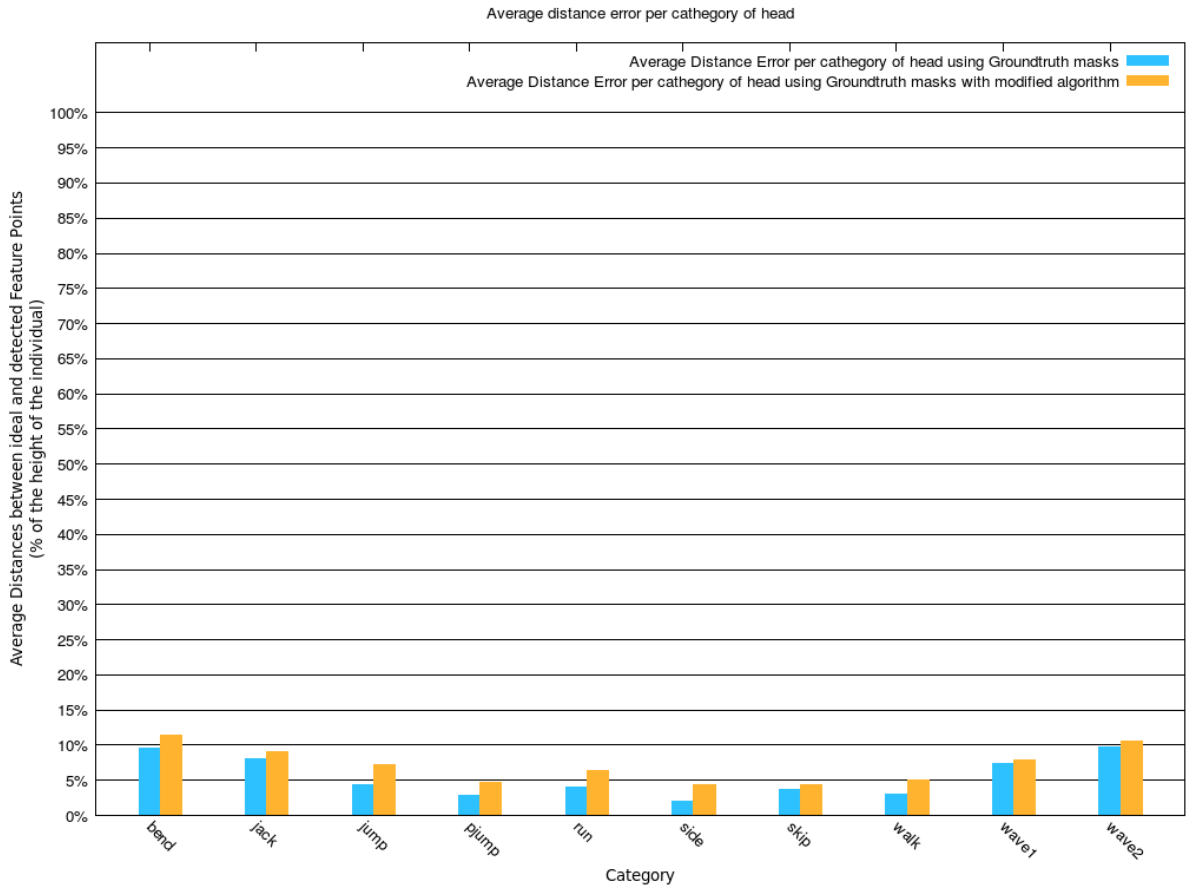


Figure A.20: Average distance to nearest point for head feature point per category using Ground-truth masks

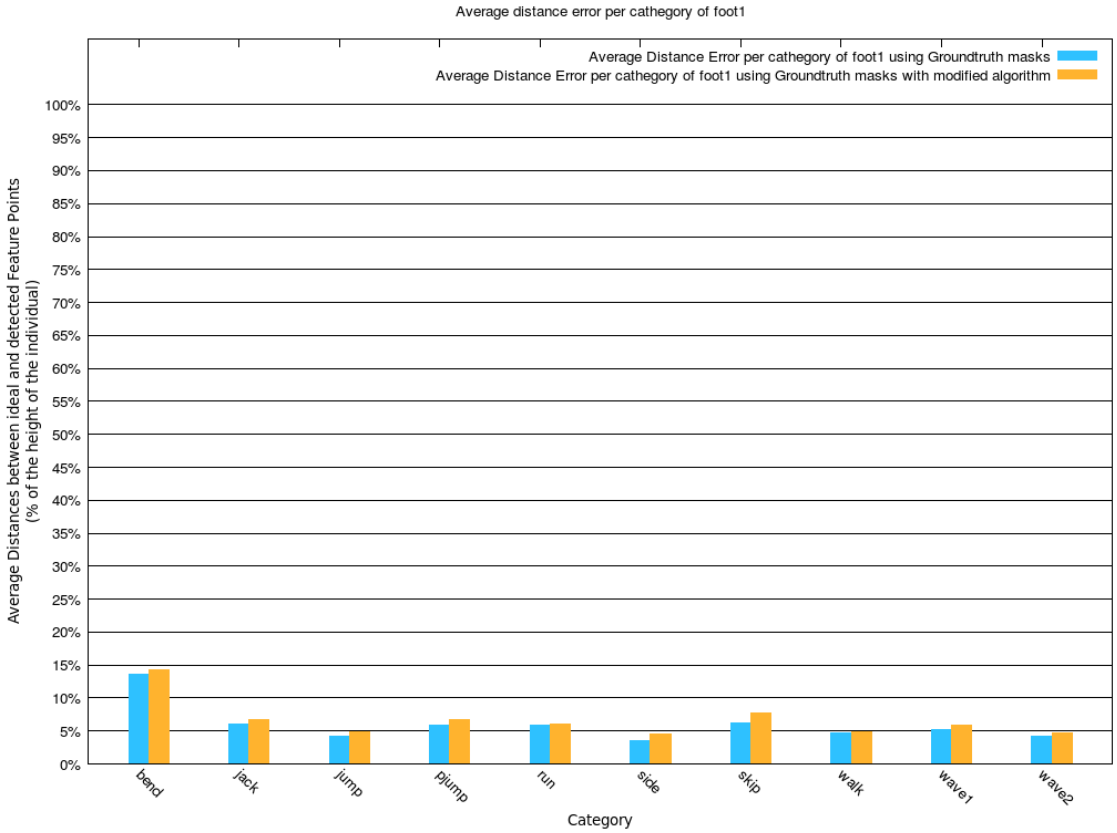


Figure A.21: Average distance to nearest point for left foot feature point per category using Ground-truth masks

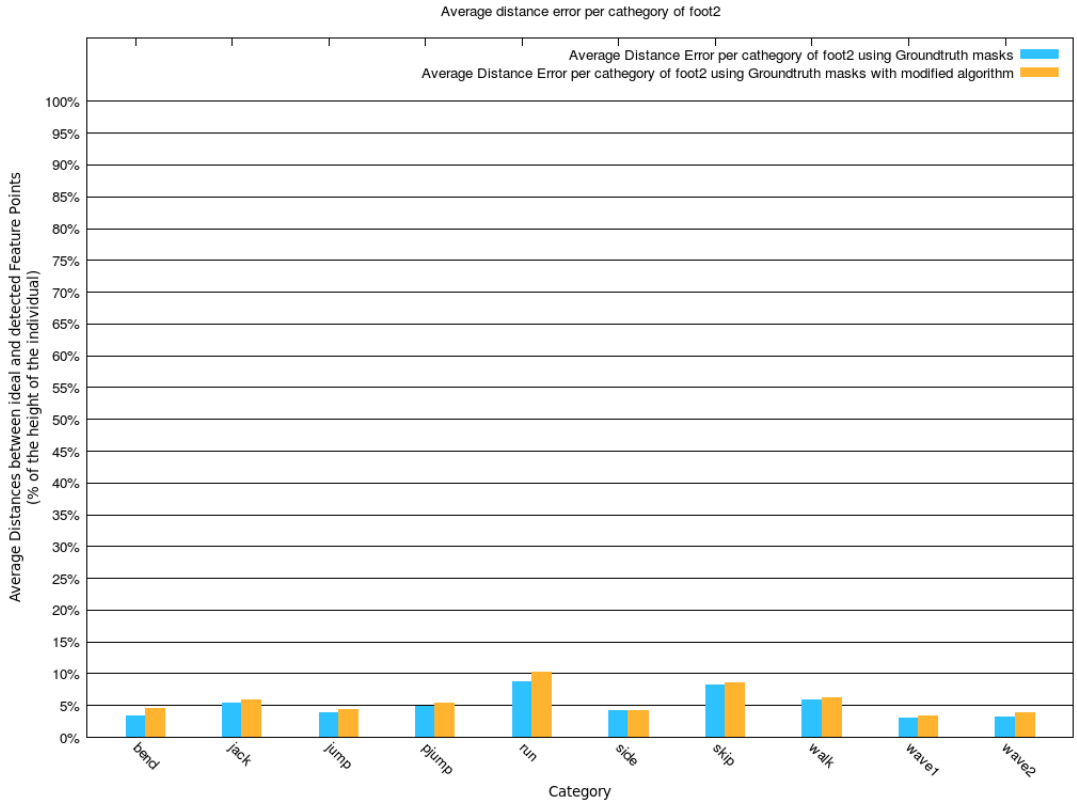


Figure A.22: Average distance to nearest point for right foot feature point per category using Ground-truth masks

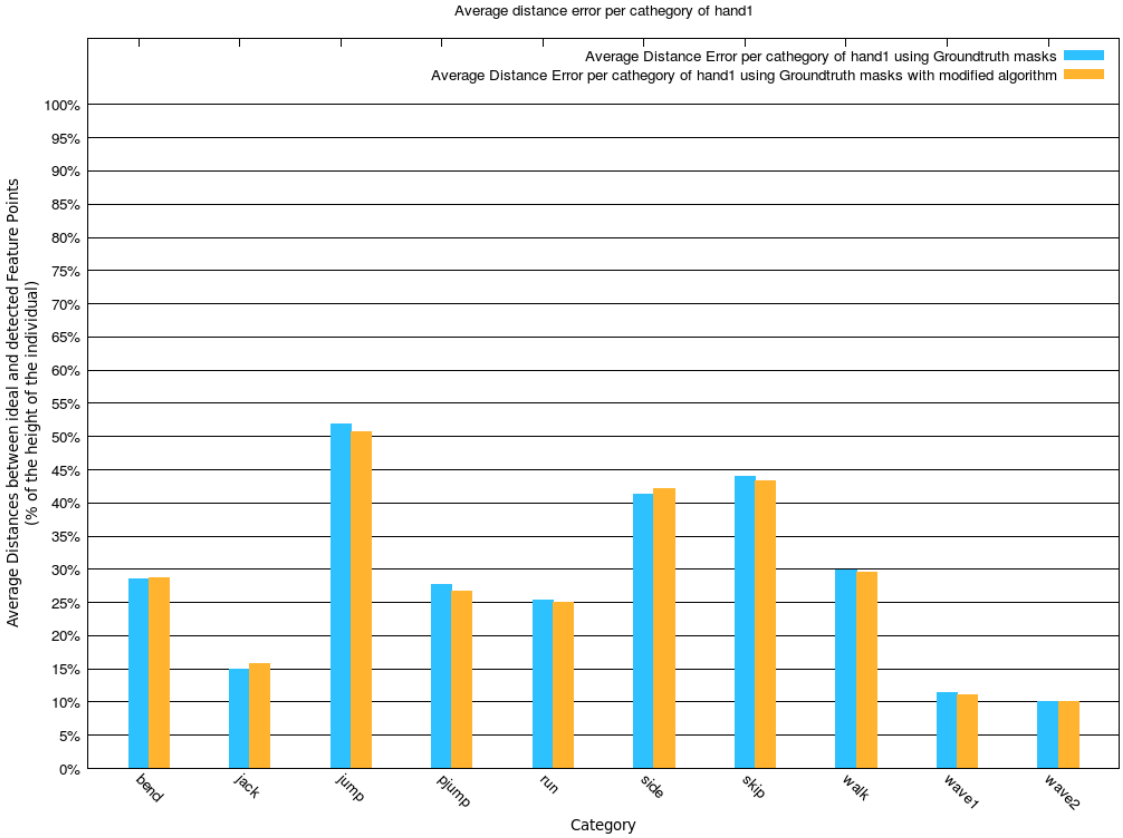


Figure A.23: Average distance to nearest point for left hand feature point per category using Ground-truth masks

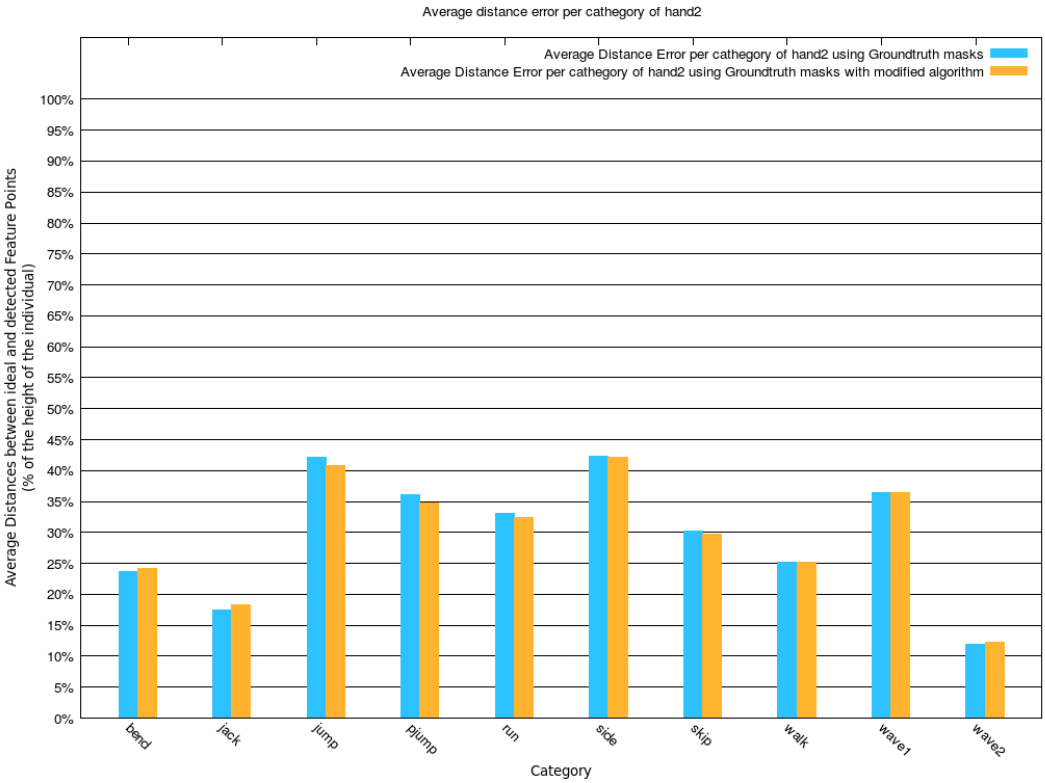


Figure A.24: Average distance to nearest point for right hand feature point per category using Ground-truth masks

A.2.3 - Angle Error

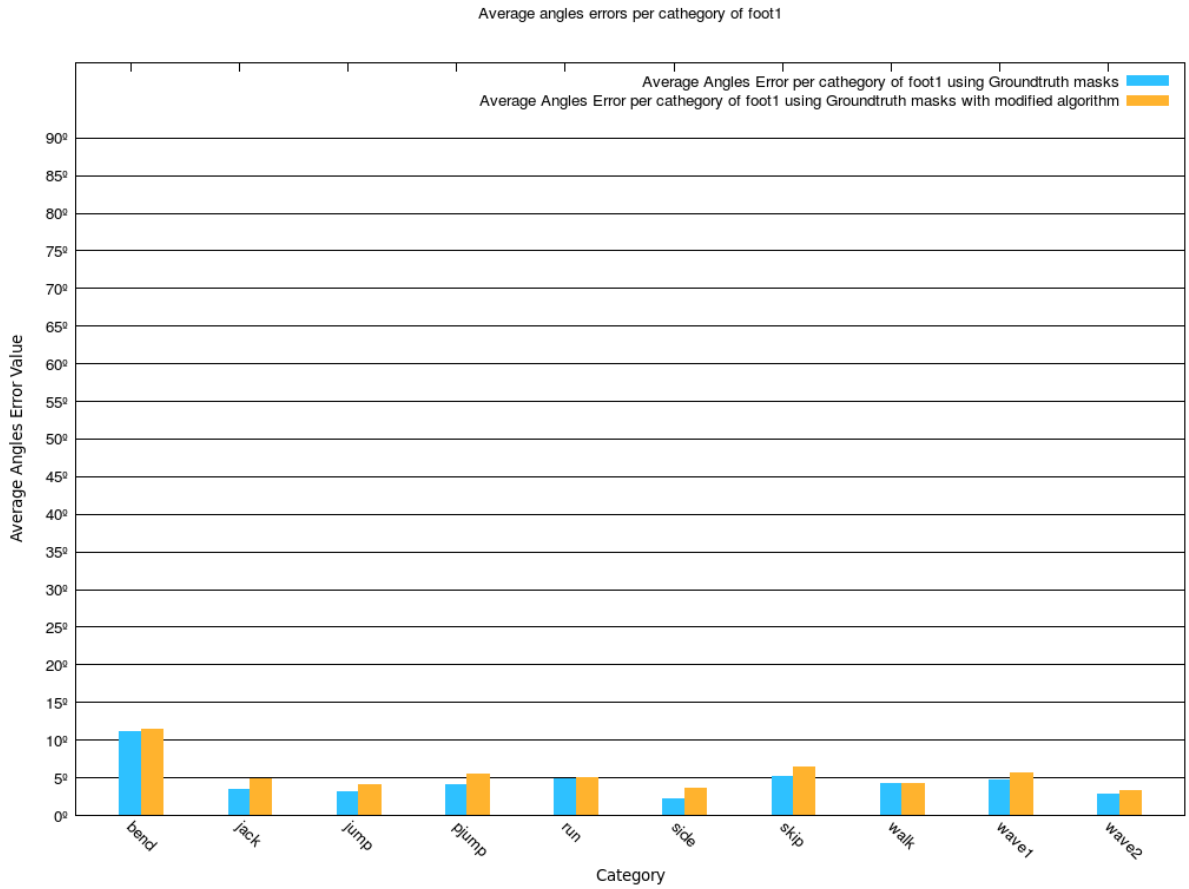


Figure A.25: Average angle error for β1 angle per category with and without post-processing using Ground-truth masks

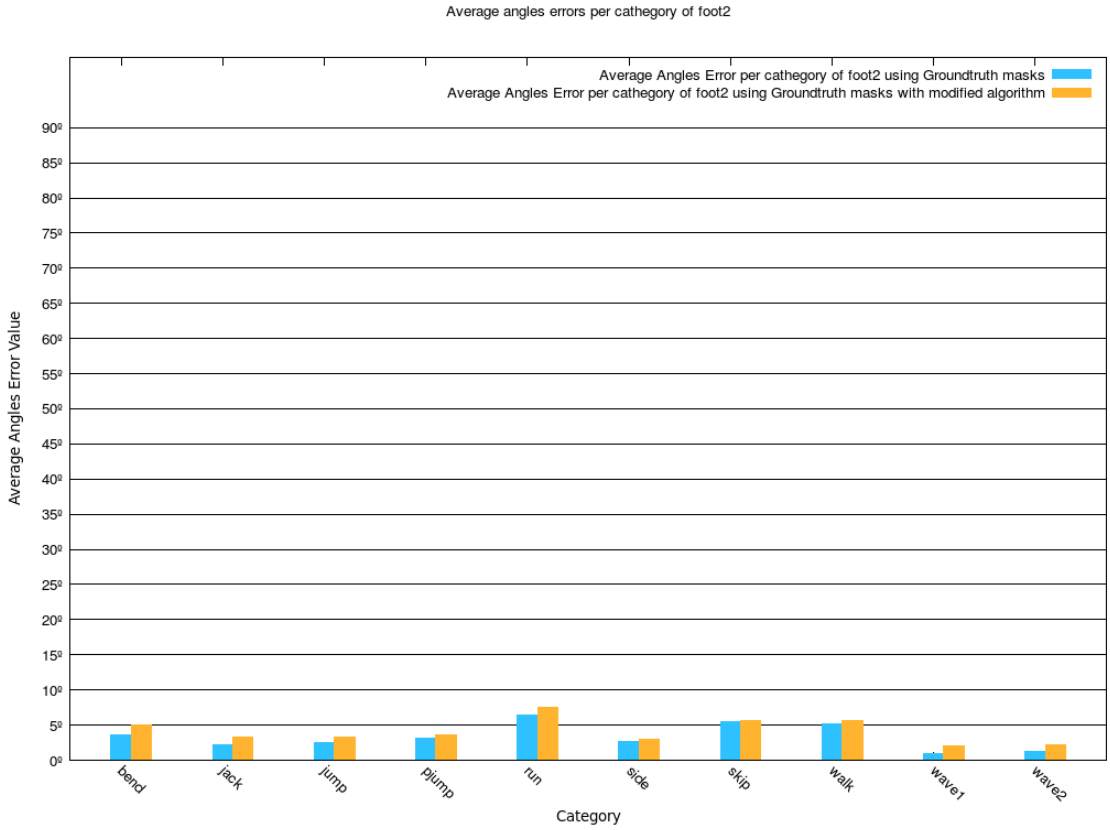


Figure A.26: Average angle error for ß2 angle per category with and without post-processing using Ground-truth masks

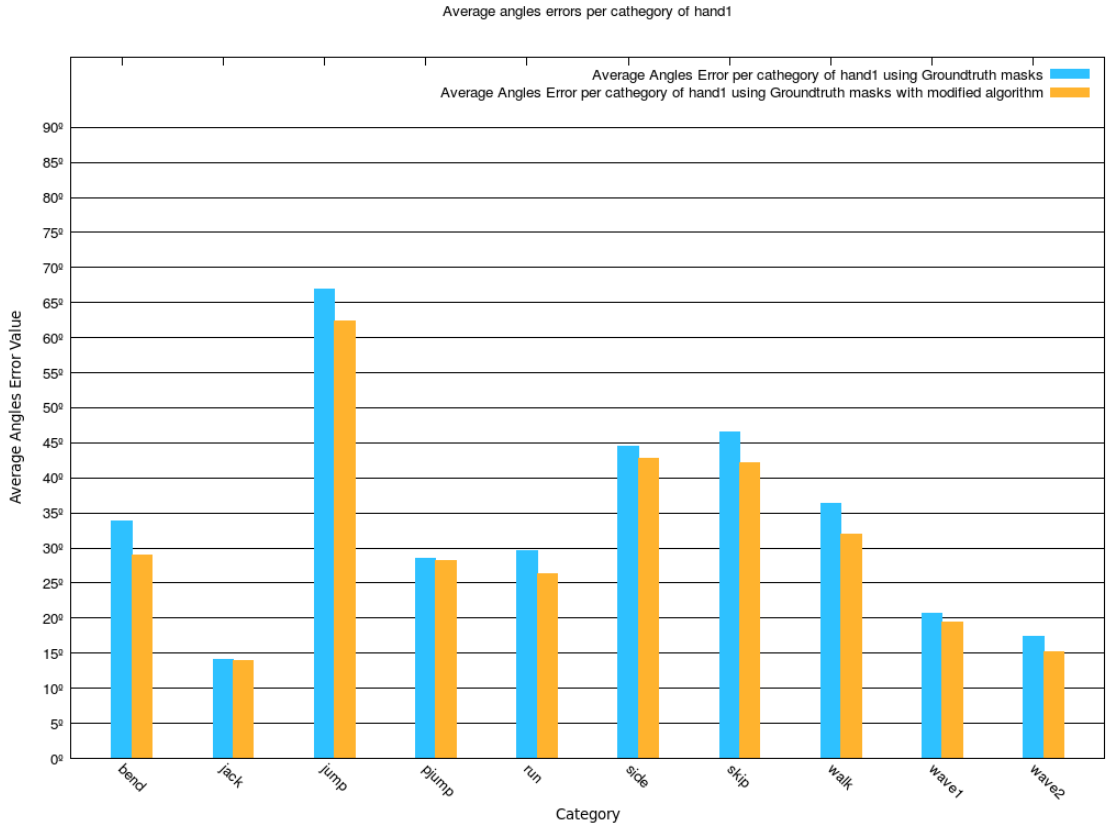


Figure A.27: Average angle error for ß3 angle per category with and without post-processing using Ground-truth masks

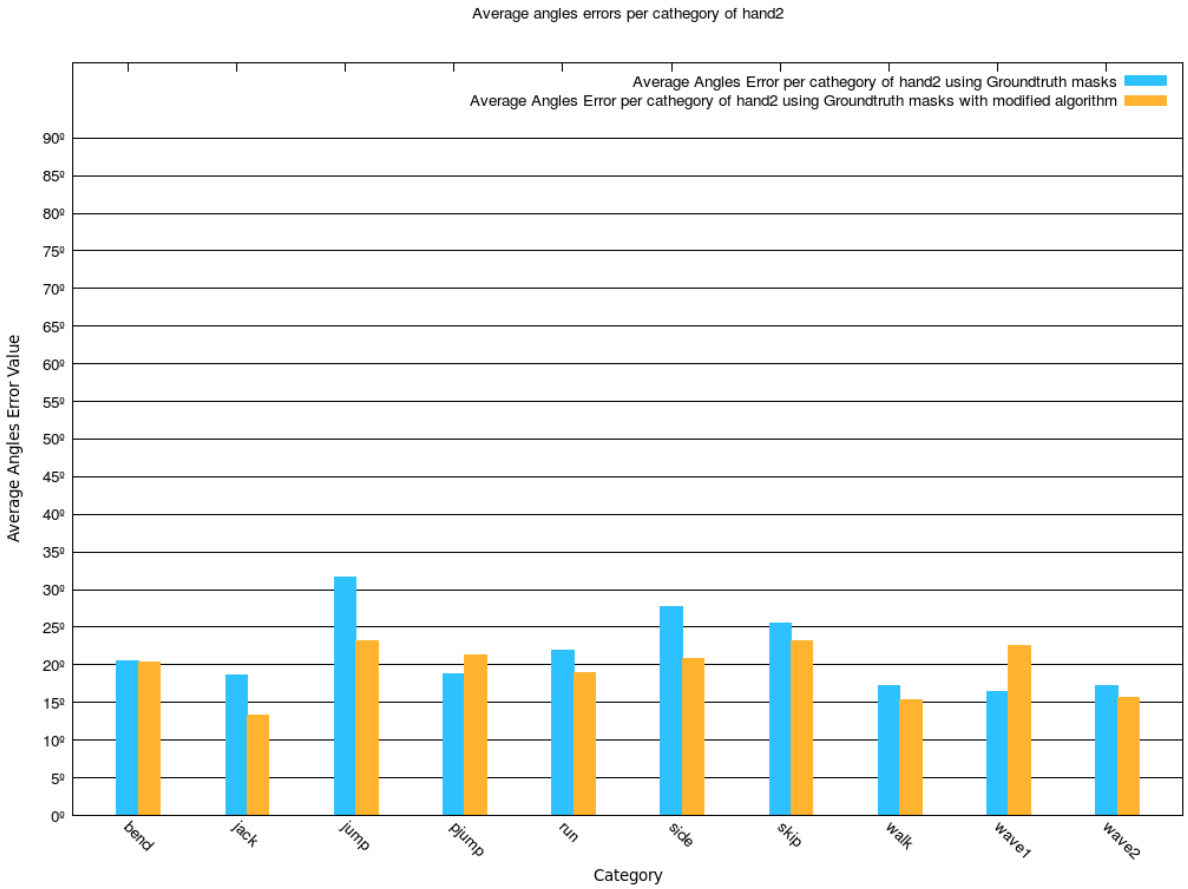


Figure A.28: Average angle error for ß4 angle per category with and without post-processing using Ground-truth masks

A.3 - Distance to reference point vs Distance to nearest point

A.3.1 - Precision

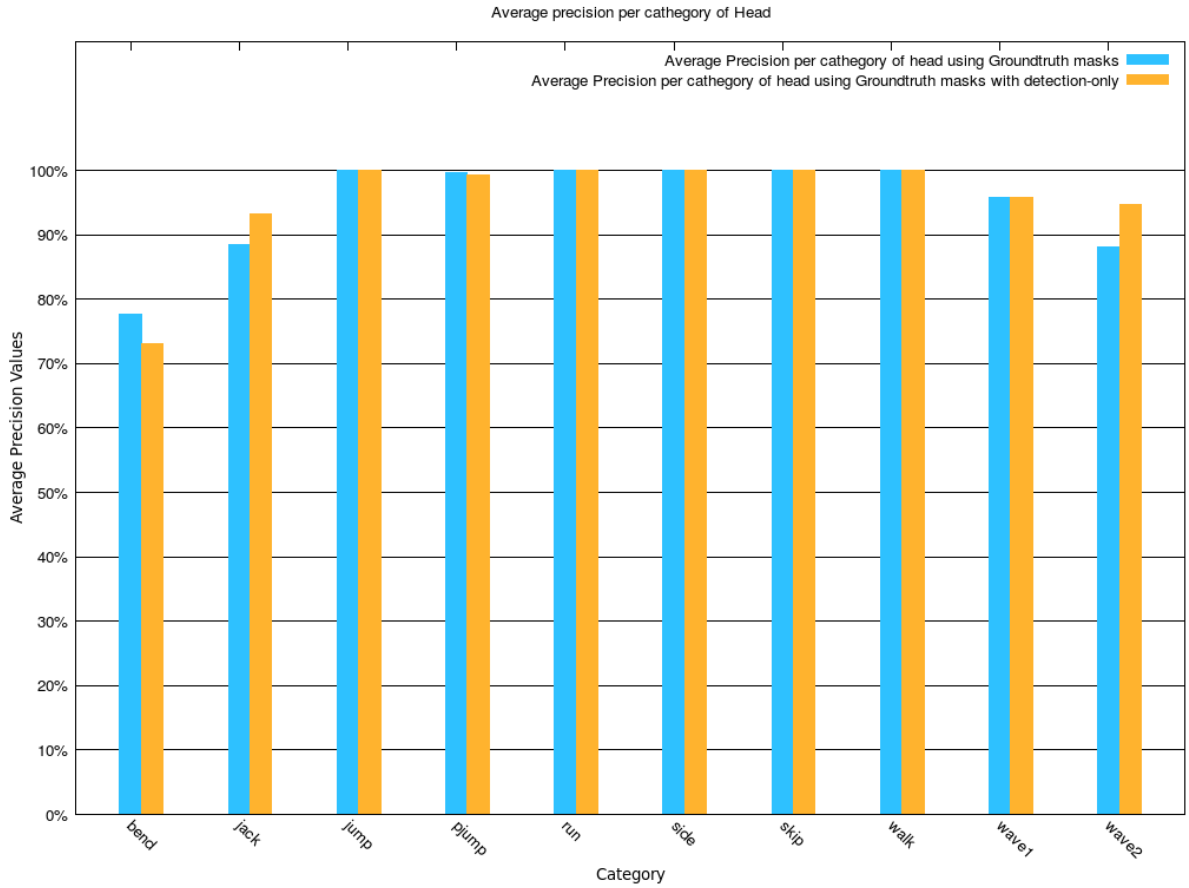


Figure A.29: Average precision for head feature point per category with distance to reference point and with distance to nearest point using Ground-truth masks

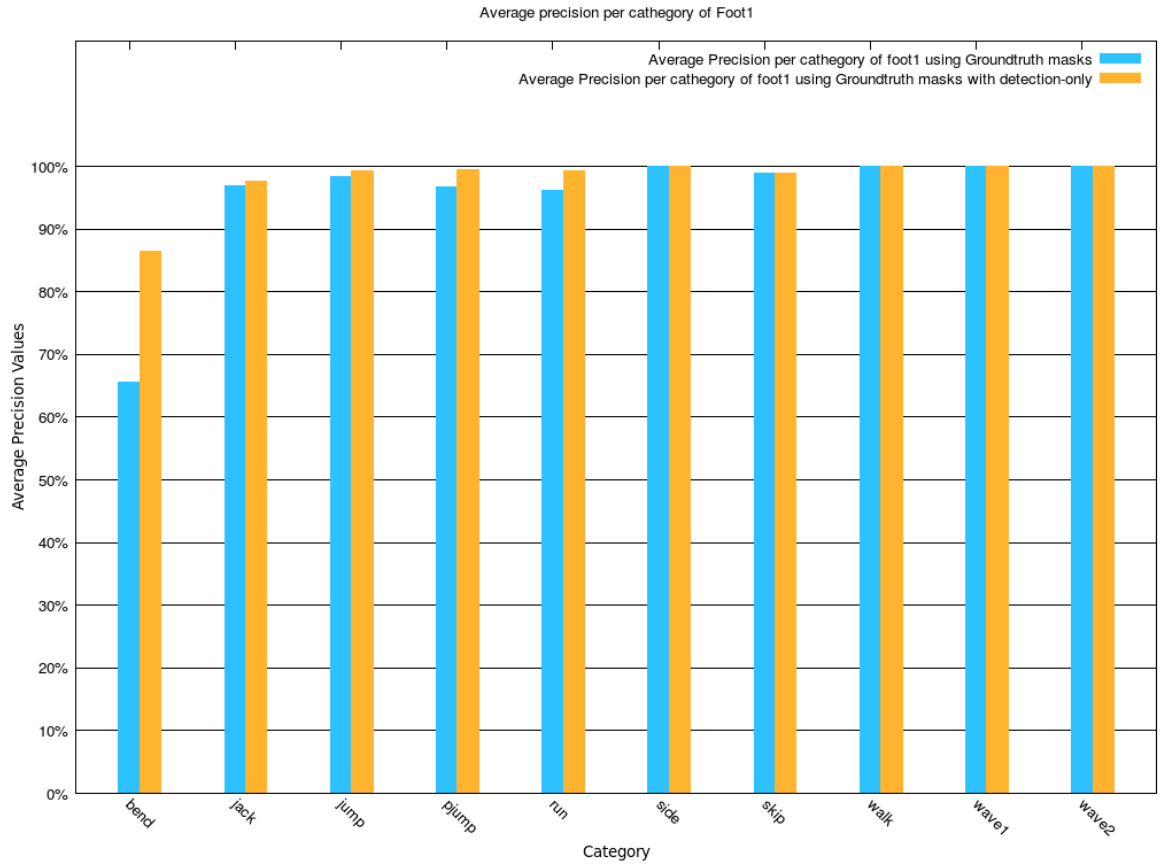


Figure A.30: Average precision for left foot feature point per category with distance to reference point and with distance to nearest point using Ground-truth masks

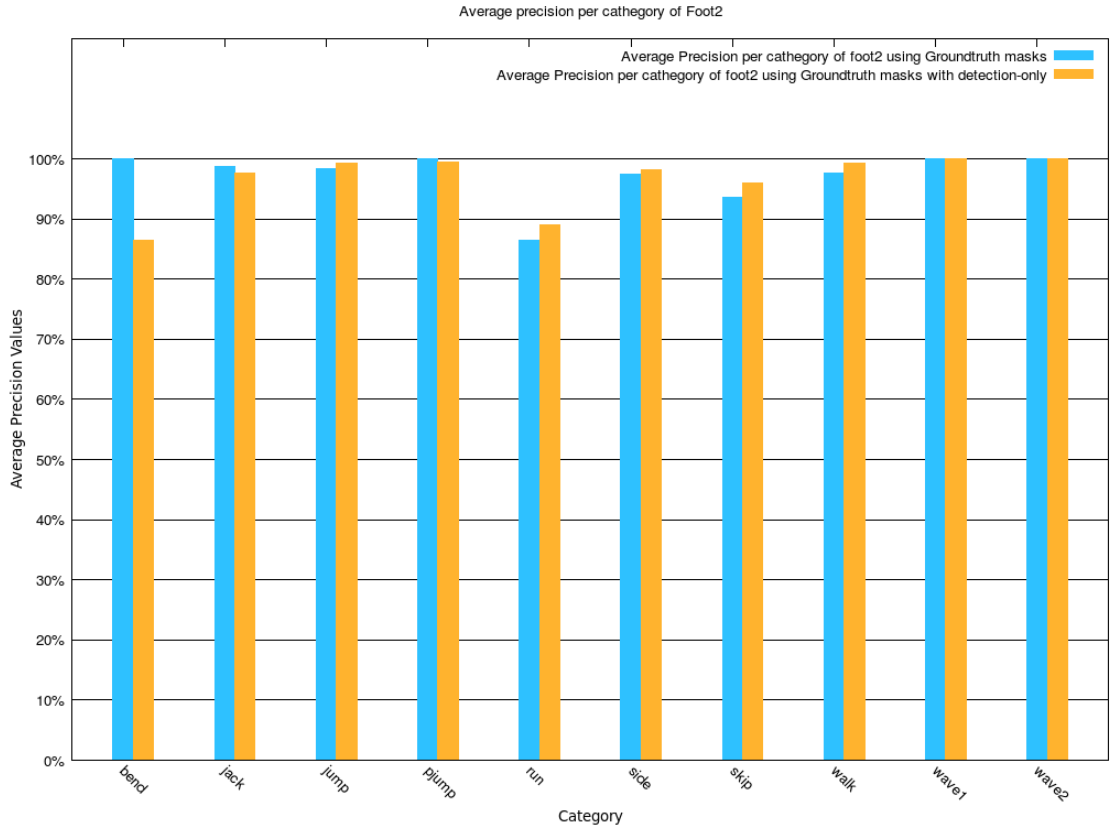
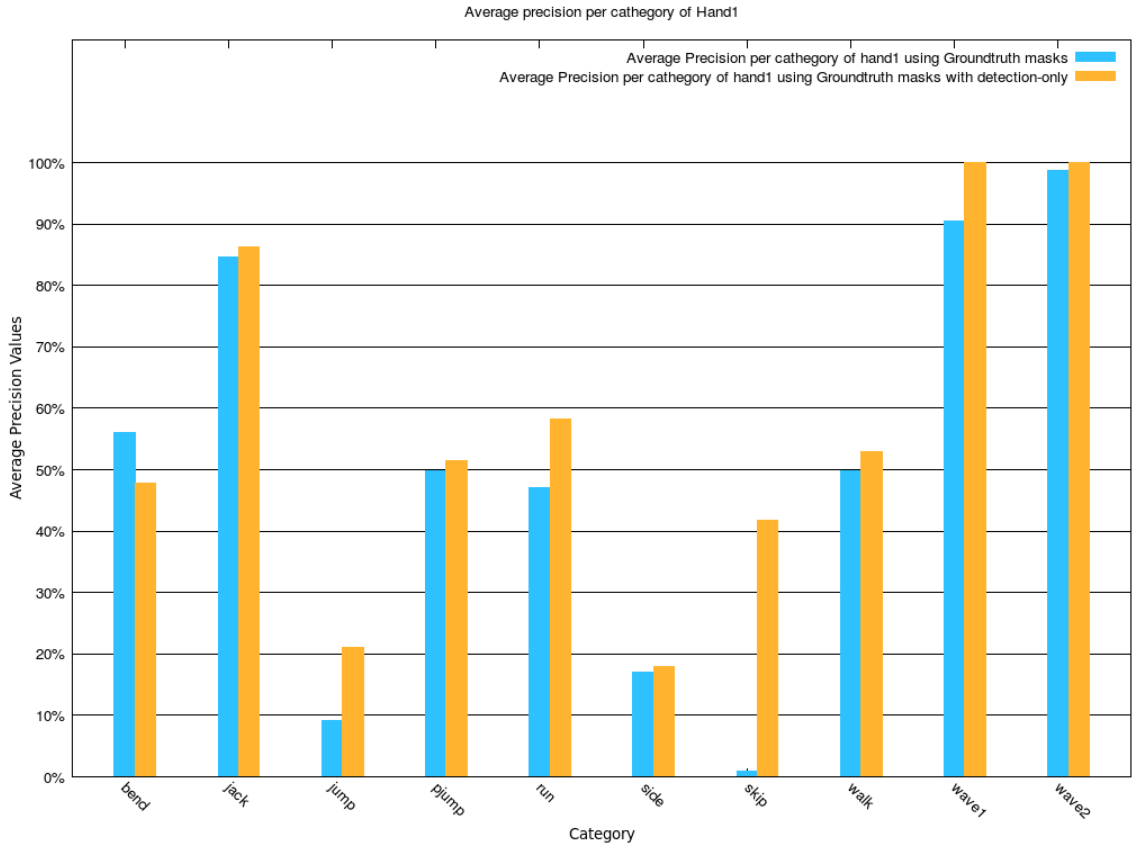


Figure A.31: Average precision for right foot feature point per category with distance to reference point and with distance to nearest point using Ground-truth masks



FigureA.32: Average precision for left hand feature point per category with distance to reference point and with distance to nearest point using Ground-truth masks

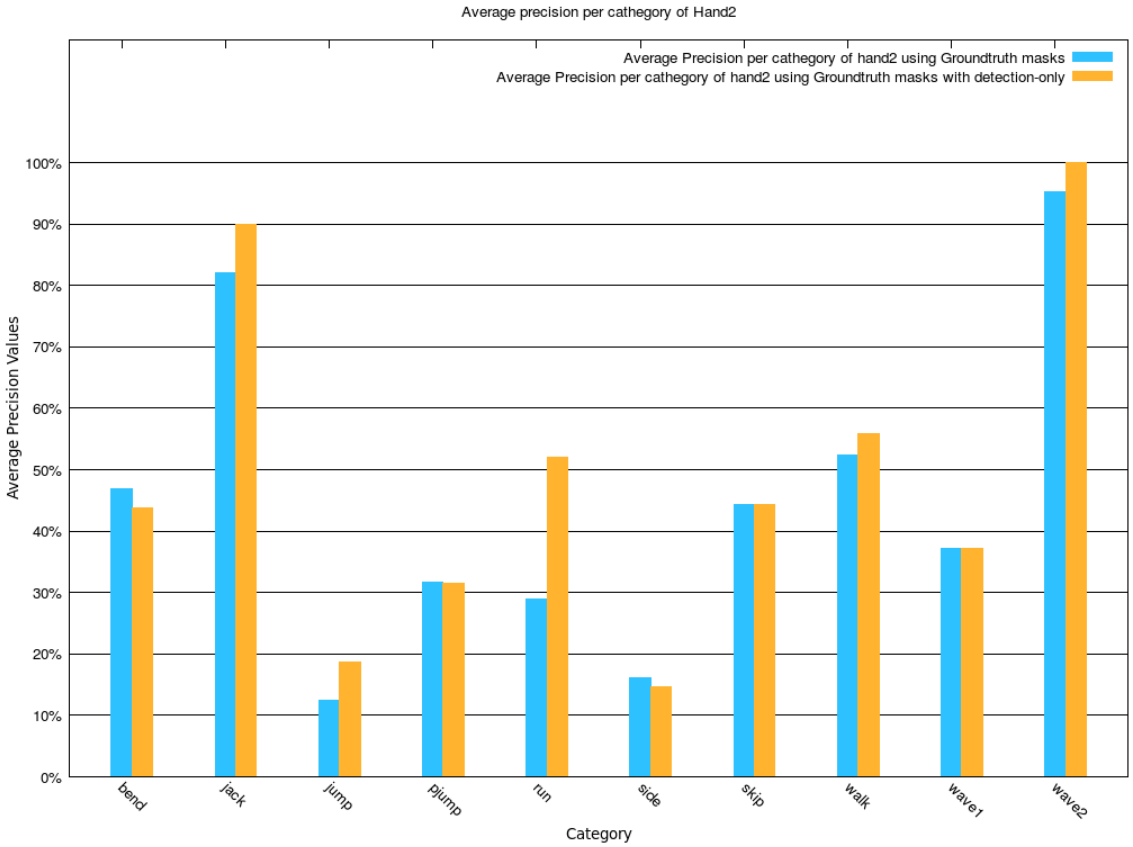


Figure A.33: Average precision for right hand feature point per category with distance to reference point and with distance to nearest point using Ground-truth masks

A.3.2 - Euclidean distance error

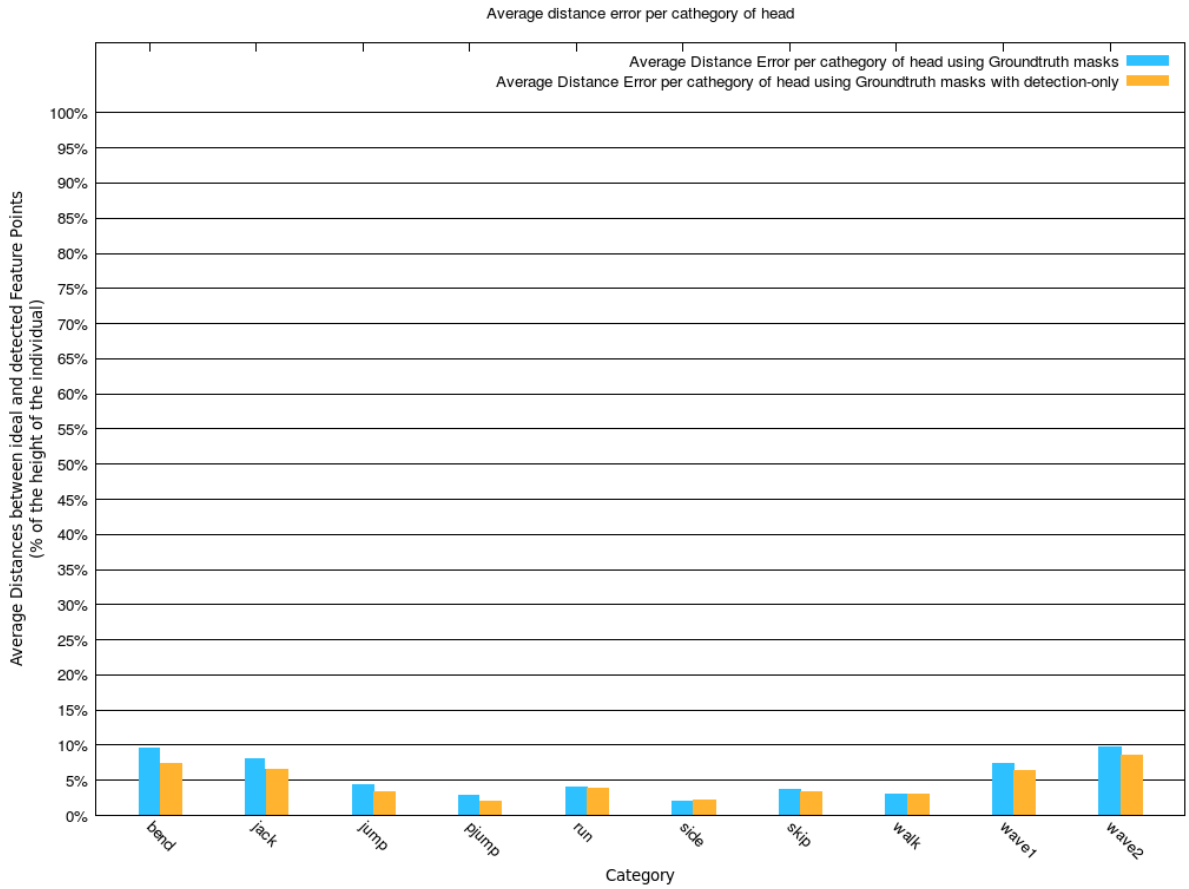


Figure A.34: Average distance to reference point and to nearest point for head feature point per category using Ground-truth masks

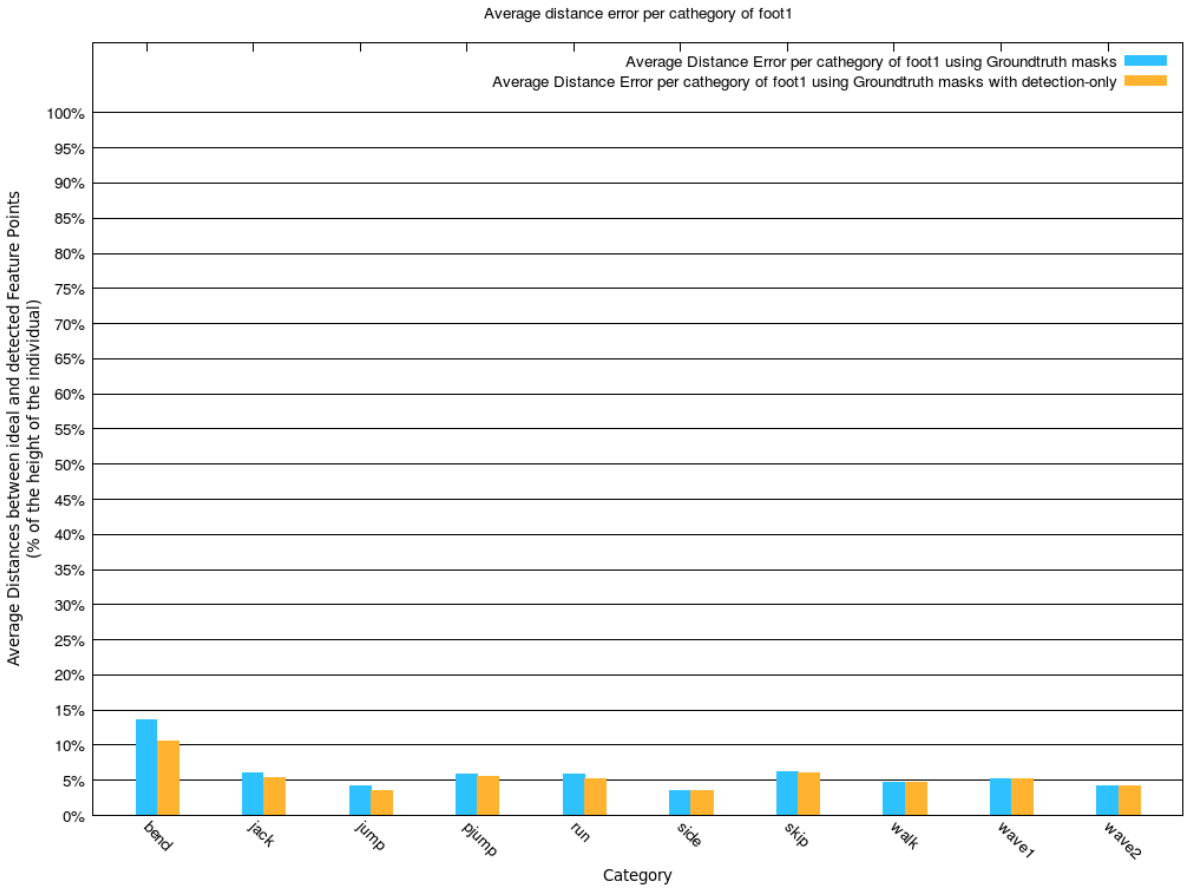


Figure A.35: Average distance to reference point and to nearest point for left foot feature point per category using Ground-truth masks

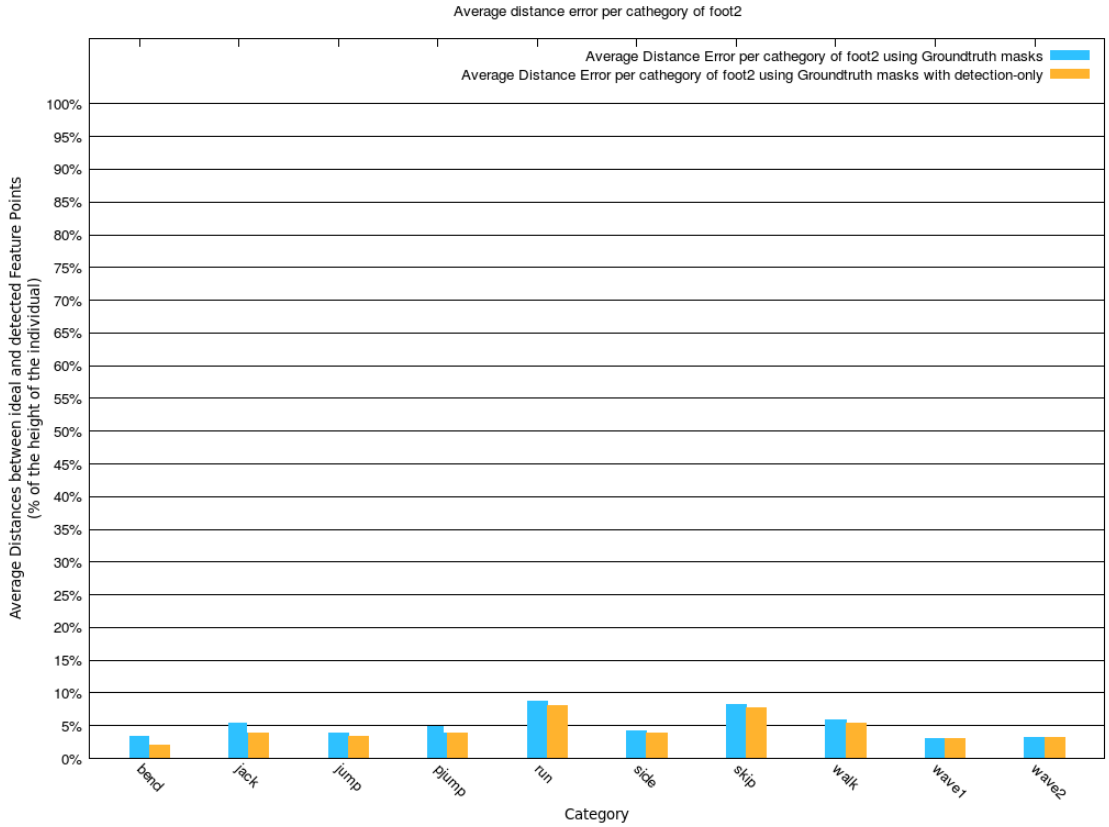


Figure A.36: Average distance to reference point and to nearest point for right foot feature point per category using Ground-truth masks

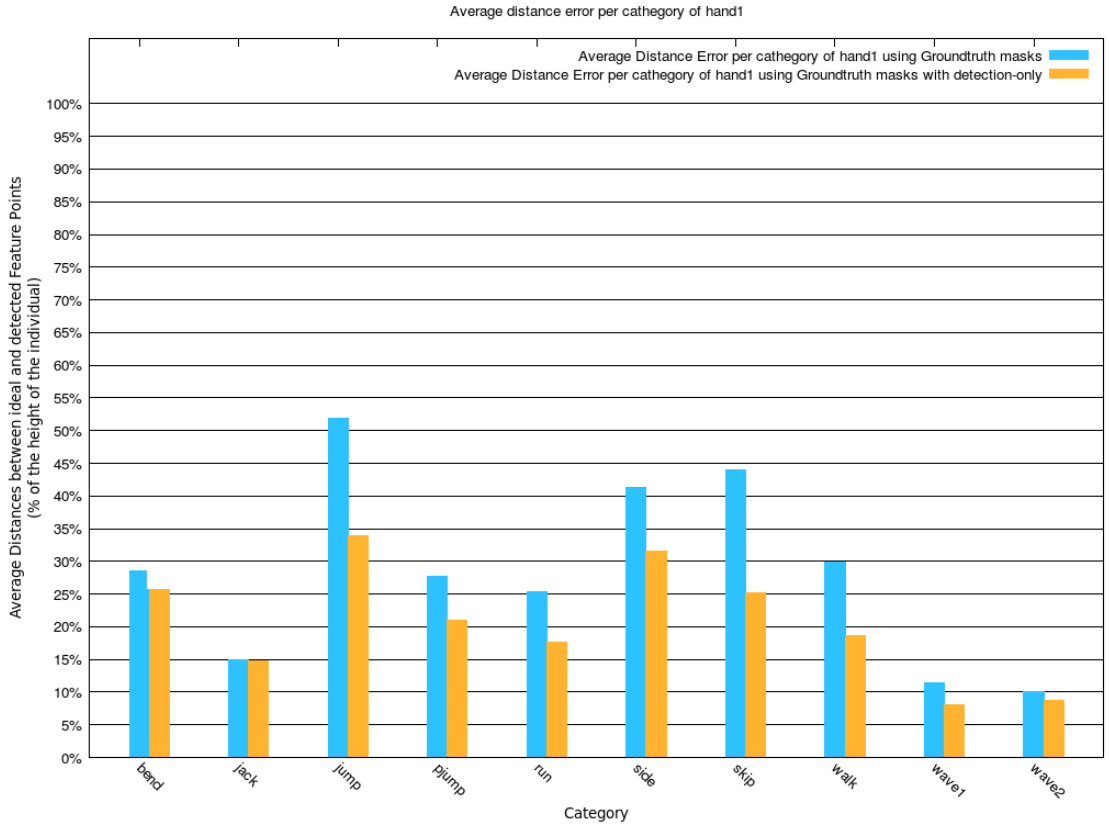


Figure A.37: Average distance to reference point and to nearest point for left hand feature point per category using Ground-truth masks

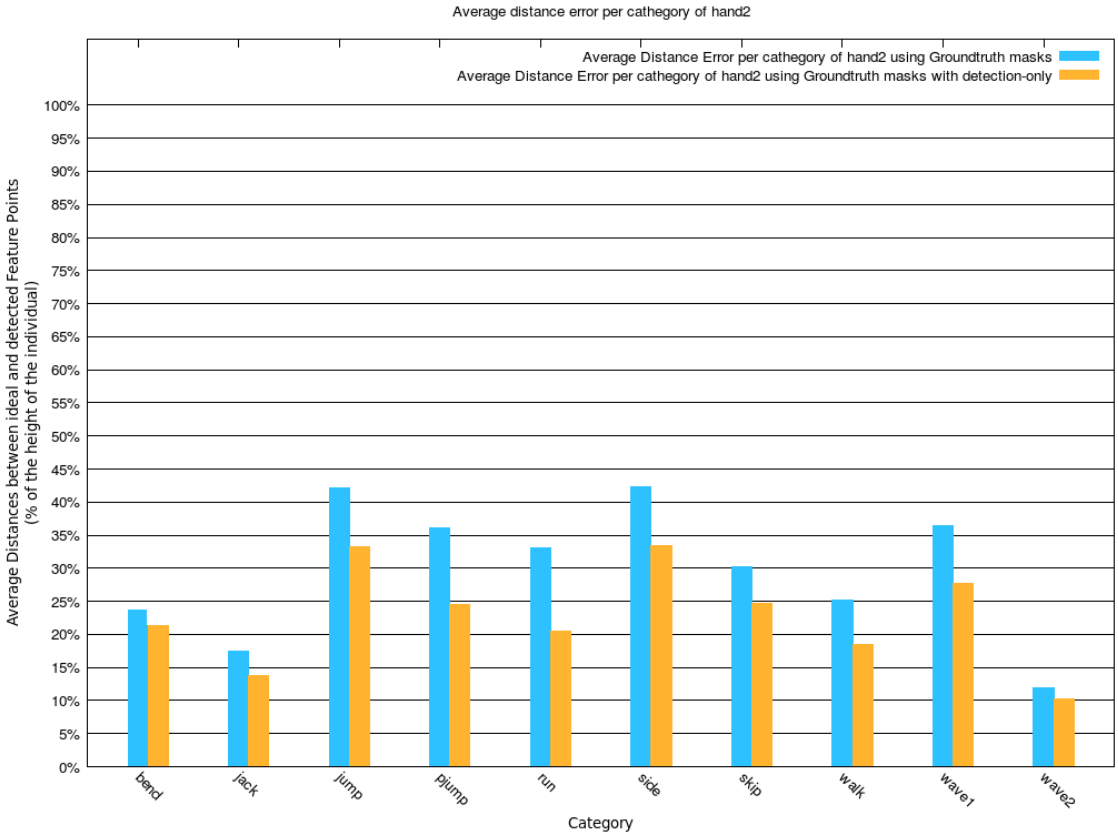


Figure A.38: Average distance to reference point and to nearest point for right hand feature point per category using Ground-truth masks

A.3.3 - Angle Error

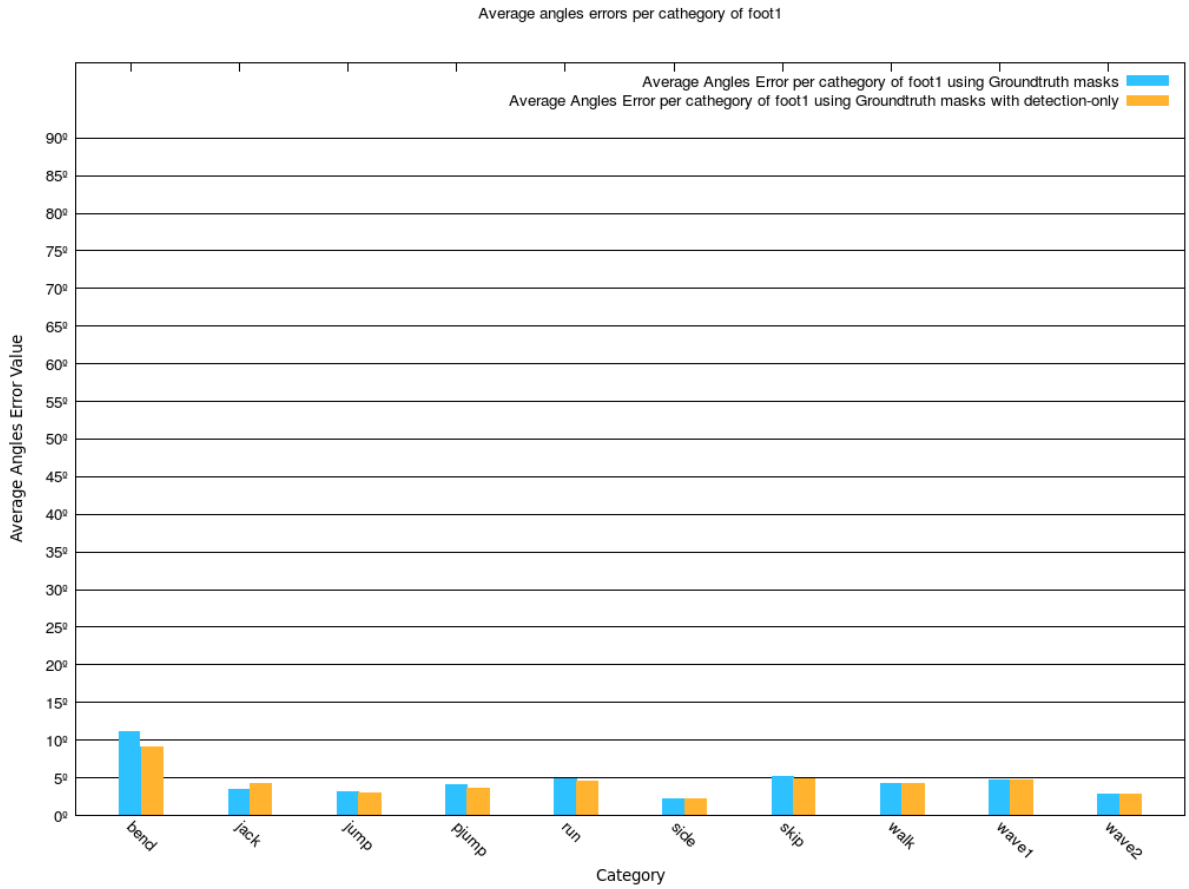


Figure A.39: Average angle error for ß1 angle per category with distance to reference point and to nearest point using Ground-truth masks

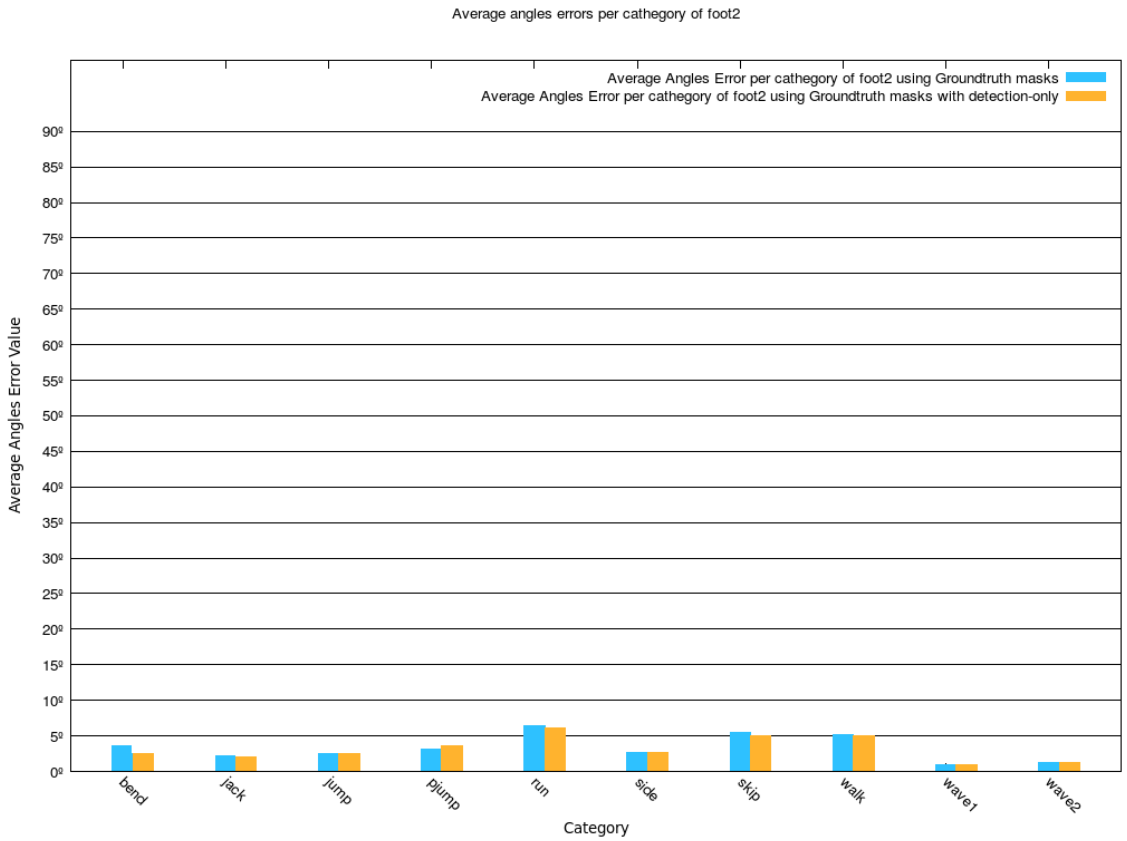


Figure A.40: Average angle error for ß2 angle per category with distance to reference point and to nearest point using Ground-truth masks

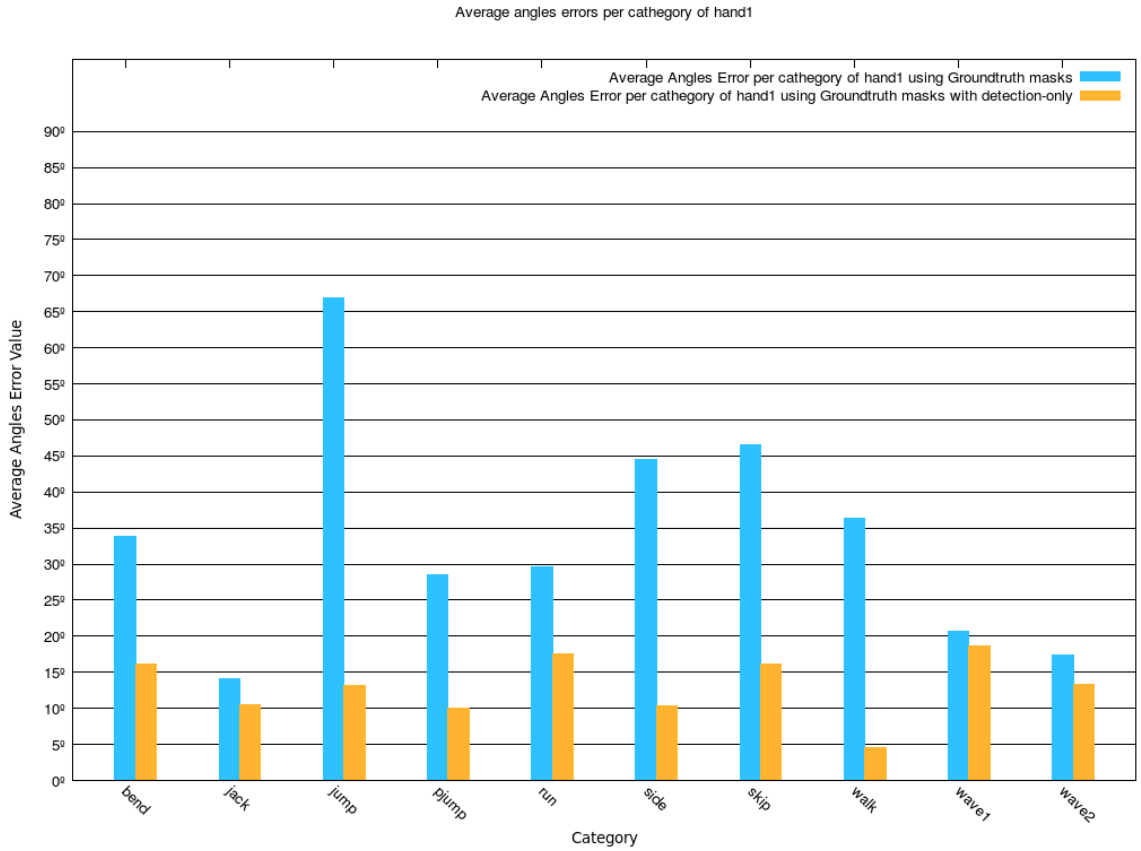


Figure A.41: Average angle error for β3 angle per category with distance to reference point and to nearest point using Ground-truth masks

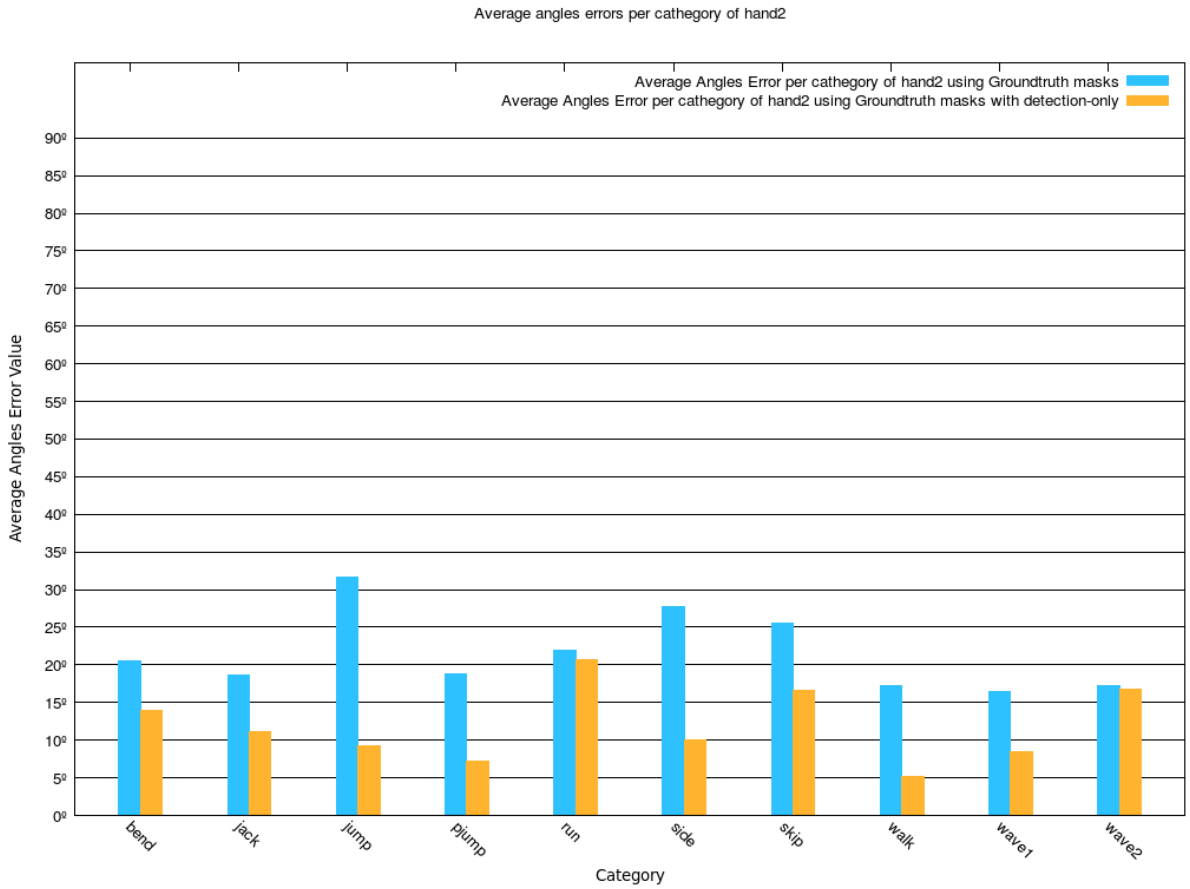


Figure A.42: Average angle error for ß4 angle per category with distance to reference point and to nearest point using Ground-truth masks