Computational algorithms for image analysis: Applications on human vocal tract and silhouette

Dissertation submitted in fulfillment of the requirements
for the degree of Doctor in Informatics Engineering by the
Faculdade de Engenharia da Universidade do Porto

Maria João Medeiros de Vasconcelos

BSc in Applied Mathematics for Technology by
Faculdade de Ciências da Universidade do Porto (2002)

MSc in Applied Statistics and Modelling by
Faculdade de Engenharia da Universidade do Porto (2006)

Supervisor

João Manuel Ribeiro da Silva Tavares
Associate Professor of the Department of Mechanical Engineering
Faculdade de Engenharia da Universidade do Porto

Co-supervisor

Miguel Fernando Paiva Velhote Correia
Auxiliary Professor of the Department of Electrical and Computer Engineering
Faculdade de Engenharia da Universidade do Porto

2015

Abstract

The present Thesis belongs to the field of Computer Vision, more specifically segmentation and analysis of objects represented in images. While Computer Vision seeks to make useful decisions about real objects and scenes based on images through the construction of artificial systems, object segmentation and analysis aims to construct capable models to efficiently characterize objects and perform segmentation in new images.

This Thesis aims to present computational algorithms for object segmentation and analysis in images suitable for application on objects such as the human vocal tract and silhouette.

The main objective for studying the vocal tract in images is to better understand the vocal tract morphology and the involved movements during speech production of the European Portuguese language. Consequently, methodologies based on statistical deformable models, namely active shape models and active appearance models, were developed to represent the vocal structures from a global perspective in magnetic resonance images.

The suggested models made it possible to obtain a realistic simulation of the vocal tract during speech production as well as efficiently perform segmentation of vocal tract in new images. Furthermore, the use of such image analysis techniques can allow for obtaining quantitative measures with higher precision and are particularly advantageous when speech therapists and imaging specialists need to analyze a large volume of data.

Regarding the human silhouette analysis, four background subtraction models were studied to segment moving silhouettes in image sequences, with different levels of complexity. In addition, and following the same methodology used for modeling the vocal tract, an active silhouette model was also developed, using information about the contour of the silhouette together with anatomical stick points and combining the shape model with its gray level profiles with the purpose of segmenting the modeled silhouettes in new images.

The results obtained from the application of the background subtraction models in four different datasets suggested that the best model depends on the complexity of the images. Moreover, the good results obtained through the use of the active silhouette model built to perform human shape segmentation in new images strongly suggests that this type of deformable model can be successfully used in this task. The main contribution accomplished regarding the modeling of human silhouettes in images is that it allows for building an active shape model that gathers the necessary information independent of the walking direction of the subject.

In conclusion, the identification and analysis of human structures are complex tasks, since their shapes are not constant and vary through time; however, techniques of Computer Vision and objects modeling can assist in their achievement as is demonstrated throughout this Thesis. To conclude, the application of the developed models in images allows realistic simulations of the human vocal tract and silhouette, making possible their competent segmentation and characterization.

Keywords: Image Processing and Analysis, Human Vocal Tract, Human Silhouette, Object Segmentation, Object Modeling, Statistical Modeling.

Sumário

A presente Tese pertence ao domínio da Visão por Computador, mais especificamente à segmentação e análise de objetos representados em imagens. Enquanto a Visão Computacional procura efetuar decisões sobre objetos reais e cenas baseado em imagens através da construção de sistemas artificiais, a segmentação e análise de imagem procura construir modelos capazes de caracterizar eficientemente objetos e efetuar segmentação em novas imagens.

Esta Tese apresenta algoritmos computacionais para segmentação e análise de imagens para aplicação em estruturas como o tracto vocal e a silhueta humana.

O objetivo do estudo do tracto vocal em imagens advém da necessidade de melhor compreender a morfologia do tracto vocal assim como os movimentos envolvidos particularmente na articulação da fala no Português Europeu. Consequentemente, foram desenvolvidas metodologias para representação global das estruturas vocais através de imagens obtidas por ressonância magnética, baseadas em modelos estatísticos deformáveis, nomeadamente modelos de forma e aparência ativa.

Os modelos desenvolvidos permitiram simular de forma realista o tracto vocal durante a articulação da fala assim como efetuar a sua segmentação em novas imagens. Para além disso, a utilização de tais técnicas de análise de imagem permitiram a obtenção de medidas quantitativas de maior exatidão e são particularmente vantajosas quando terapeutas da fala e imagiologistas necessitam de analisar grandes volumes de dados.

Em relação à análise da silhueta humana, foram estudados quatro métodos de subtração de fundos para segmentação de objetos em movimento em sequências de imagens com diferentes níveis de complexidade. Foram ainda desenvolvidos modelos de silhueta ativos, que utilizam a informação do contorno da silhueta conjuntamente com pontos anatómicos, com o objetivo de segmentar as silhuetas modeladas em novas imagens, através da combinação do modelo da forma com os seus perfis de cinzento.

Os resultados obtidos pela aplicação dos métodos de subtração de fundos em quatro bases de imagens distintas sugerem que o modelo ideal depende fortemente da complexidade da imagem em causa. Os bons resultados obtidos pela aplicação dos modelos de silhueta ativos para segmentação de silhuetas humanas em novas imagens demostram que este tipo de modelos deformáveis pode ser utilizado nesta tarefa. O principal resultado obtido em relação à modelação da silhueta humana através de imagens concerne ao facto do modelo sugerido permitir construir um modelo da forma ativo que reúne a informação da silhueta independentemente da direção do movimento do sujeito.

Em conclusão, a análise automática do tracto vocal e da silhueta humana em imagens são tarefas complexas, pois estas estruturas apresentam formas complexas bem como variáveis; no entanto, a Visão por Computador e a modelação de objetos podem ser utilizadas de forma a auxiliar tais tarefas, como se demonstra ao longo desta Tese. Assim, os modelos desenvolvidos permitem simular a forma do tracto vocal e a silhueta humana assim como efetuar com sucesso a segmentação e caracterização de tais estruturas em novas imagens.

Palavras-Chave: Processamento e Análise de Imagem, Tracto Vocal Humano, Silhueta Humana, Segmentação de Objetos, Modelação de Objetos, Modelação Estatística.

Acknowledgment

I would like to express my gratitude to my supervisor, Prof. João Manuel R. S. Tavares, for his continuous support throughout the project and encouragement for scientific production. Prof. Tavares incentive made me choose for the research career.

I would also like to thank to my co-supervisor, Prof. Miguel F. P. V. Correia for his availability whenever I needed.

To Sandra Ventura, colleague in this journey, by the collaboration in the work related with vocal tract modeling and opportunity to develop such models.

To my colleagues and friends from L304 especially Carla and Andreia for their presence and support during these years.

To my friends Isabel, Liliana, Camila, Rita, Octávio and Cláudia for the encouragement.

To my parents, my brother, to Luís and my family for the patience and for believing in me. To my grandmother that never understood why I didn't finish my studies like everybody else, this is for you.

This work was supported by the PhD grant SFRH/28817/2006 from Fundação para a Ciência e Tecnologia (FCT), in Portugal.

Contents

List of Figures

List of Tables

Acronyms

1.5T – 1.5 Tesla

3.0T – 3.0 Tesla

AAM – Active Appearance Model

ACL – Anterior Cruciate Ligament

ASM – Active Shape Model

CASIA – Institute of Automation Chinese Academy of Sciences

CMU – Carnegie Mellon University

CT – Computed Tomography

EMA – Electromagnetic Articulography

EP – European Portuguese

FIPM – Football Interaction and Process Model

FN – False Negatives

FP – False Positives

GMM – Gaussian Mixture Model

HBA – Human Behavior Analysis

HBU – Human Behavior Understanding

HOG – Histogram of Orientated Gradients

ICP – Iterative Closest Point

IPA – International Phonetic Alphabet

KTH – Kungliga Tekniska Hogskolan

MICA – Modified Independent Component Analysis

MoCap – Motion Capture

MPI08 – Indoor Motion Capture Dataset

MR – Magnetic Resonance

MRI – Magnetic Resonance Imaging

MuHAVi – Multicamera Human Action Video Dataset

OCR – Optical Character Recognition

P – Precision

PCA – Principal Component Analysis

PDM – Point Distribution Model

R – Recall

SIFT – Scale Invariant Feature Transformation

TN – True Negatives

TP – True Positives

UMPM – Utrecht Multi-Person Motion benchmark

1

Introduction

The domain of Computer Vision seeks to make useful decisions about real objects and scenes based on images. It is a multidisciplinary domain of science and technology that depends on the information taken from images for designing artificial systems that aim to simulate human vision [Ballard et al. 1982].

The evolution of this domain is strongly influenced by the need for identifying, tracking and analyzing objects in an image or a sequence of images. In order to do this, it is necessary to perform tasks such as object modeling, segmentation, tracking and analysis [Szeliski 2010]. Segmentation and analysis of objects represented in images are two of the more studied and developed tasks in computer vision, wherein various methodologies have been used to build models capable of efficiently characterizing objects.

In this Thesis, particular attention is given to the use of deformable models in image analysis, which include segmentation techniques such as template matching, active contours, deformable templates, statistical methods, level set methods and physical methods [Zhang 2001; Tavares et al. 2009].

Template matching consists of comparing the template images with the new image and searching for similarities between the two images [Schalkoff 1989; Carvalho et al. 2005]. For example, in [Carvalho et al. 2005] a template of a human eye is used to segment the eye into new images through image correlation.

The use of deformable models in image analysis and interpretation was first introduced by [Kass et al. 1988], in which *snakes* are presented. A *snake* is an active contour that adjusts to a given object through a combination of internal and external forces, where the internal forces translate the flexibility and stretch, and the point at which the external forces pull the contour towards relevant areas of the image. The adjustments of the active contour are stopped when a minimal energy state is reached, typically when it finds the object border.

Other types of deformable models are deformable templates, which use templates described by parametric functions [Carvalho et al. 2007]. The geometrical templates are defined by parameters which describe the expected geometrical shape of the object and interact dynamically with the image during the segmentation process, similarly as with *snakes*. For instance, in [Yuille et al. 1992], the authors build a model to detect eyes in images, where the eye is represented by a circle describing the iris, two parabolic curves describing eyelids and also the intensity of these regions. The combination of all these characteristics typically translates into a model with high number of parameters, making the construction of deformable templates complex whenever a new object type needs to be modeled [Tien et al. 2000].

Statistical models are also included in the category of deformable models. An example of such modeling technique is given in [Carvalho et al. 2007] to identify skin areas in an image. For this, sample images of skin are used to build a statistical model for posterior skin segmentation that indicates the probability of the pixels of the new image to be associated with human skin. Another example of statistical modeling are the Point Distribution Models (PDMs) that were initially proposed by [Cootes, Taylor, et al. 1992] to model objects based on its statistical analysis. These models are obtained through the analysis of the statistics of the coordinates of the landmarks that represent the deformable object under study: after aligning the object shapes, a principal component analysis is made and the mean shape of the object and the main modes of its variation are obtained.

Active Shape Models [Cootes and Taylor 1992] and Active Appearance Models [Cootes et al. 1998] use PDMs to segment and recognize the modeled objects in new images. Both models use a combination of the statistical shape model with the gray levels of the object's landmarks.

The idea of considering physical constraints in object modeling has been suggested and used by several authors. In [Pentland et al. 1991], the authors present physical-based solution for modeling objects. The approach is based on the finite element method and parametric solid modeling using implicit functions. Also in [Gonçalves et al. 2009] a physical approach based on the finite element method is used to segment an object and simulate its deformation. For this, an initial contour is manually defined that automatically evolves until it converges to the border of the desired object.

Another possibility to perform image segmentation of objects is to use level set methods, introduced by [Osher et al. 1988]. The main idea behind these methods is to represent the moving contour using a signed function whose zero corresponds to the actual contour. Then, by tracking the zero level set of the function adopted in the modeling it is possible to derive a similar flow for the implicit surface. A survey of algorithms that combine statistical techniques with level set methods can be found in [Cremers et al. 2007]. For example, in [Ma et al. 2013] a level set based algorithm is proposed to reconstruct the 3D shape of the bladder using cross-sectional boundaries in magnetic resonance images.

The analysis of objects in images has been encouraged by the improvement of human/machine interaction in several applications, covering fields from industrial inspection, optical character recognition (OCR), medical imaging, surveillance or fingerprint recognition and biometrics. In industrial inspection it is used mostly for quality control purposes or defect recognition [Agin 1980; Klein et al. 1994; Campos et al. 2010]. Regarding OCR, examples of applications include reading handwritten postal codes on letters and automatic number plate recognition [Matan et al. 1992; Fahmy 1994; Volna et al. 2013]. Applications in medical imaging include performing image registration [Ayache 1998; Damas et al. 2011]. Computer Vision can also aid in the designing of surveillance systems for detecting and monitoring intruders or analyzing highway traffic [Sage et al. 1999; Norouznezhad et al. 2008]. Fingerprint recognition and biometrics has been used extensively for automatic access authentication and forensic applications [Junta Doi et al. 2004; Garibotto 2009; Nadipally et al. 2013].

1.1. Objectives

The subject of object analysis in Computer Vision has been developing in recent decades; especially in the domains of the analysis of objects in medical images and the human body, two of its most active fields.

The identification and analysis of human structures are complex tasks, since their shapes are not constant and vary through time; however, techniques of Computer Vision and objects modeling can assist in their achievement as one aim to demonstrate throughout this Thesis.

This Thesis is dedicated to developing computational algorithms for object segmentation and analysis in images. The human vocal tract and silhouettes were the objects selected to be analyzed in this Thesis. Hence, the objectives defined in this project included:

- Review the existing algorithms for image analysis used to characterize and segment the human vocal tract and the human silhouette;
- Analyze the need to develop methods for application in objects such as the vocal tract and human silhouette and define suitable clues that can be used to enhance the segmentation;
- Develop new computational algorithms for characterizing such objects in images, particularly highlighting techniques based on the modeling of the geometrical shape of the object as well as its behavior;
- Test the developed algorithms for segmenting the objects in new images and analyze the segmentation results, both qualitatively and quantitatively; compare the algorithms with existing ones and find the positive aspects and drawbacks of these algorithms.

1.2. Organization of the Thesis

After this introduction, in Chapter 2 the state-of-the-art of computational algorithms for image analysis used for studying the human vocal tract and human motion analysis are reviewed. A brief description and explanation are provided for the vocal tract anatomy and the most common imaging techniques used to acquire images of the complete vocal tract. The most promising methods used to represent its shape, including a summary of speech production studies in various languages available at the moment are also provided as well as the importance of vocal tract modeling. Regarding application on human motion analysis, the most important related research is presented, together with a description of the methodologies used for human detection, tracking and understanding; existing applications and existing datasets are also referred to and challenges are pointed out.

Chapter 3 presents the developed methodology to segment the shape of the vocal tract in new images for speech production assessment. A description is provided for the sounds of European Portuguese language and Point Distribution Models, Active Shape Models and Active Appearance Models. The image datasets and the Magnetic Resonance Imaging protocols used to build the models are also described. The segmentation results of the various active shape models developed for the study of the shape and appearance of the vocal tract shape are also presented and discussed, as well as an example of their application to actual studies.

The methodologies developed to segment silhouettes from images sequences are reported in Chapter 4. Four different background subtraction models are addressed in this chapter and an active silhouette model built is presented. Different image sequences were used for testing the developed methods and quantitative results are presented and discussed.

Finally, in Chapter 5 the main conclusions are drawn and suggestions for future research are given.

1.3. Contributions

Throughout this project, computational algorithms for image analysis were developed for application on human vocal tract and silhouettes. During this period, one book chapter was published and another was accepted for publication; six papers derived from the Thesis have been published in peer-reviewed journals; additionally, four papers and eight abstracts have been included in conference proceedings. In addition one symposium was organized during this Thesis.

The main contributions of this Thesis can be summarized as the following:

- A comprehensive review of the current computational algorithms for image analysis that have been used for the study of the human vocal tract during speech production and for the study of human motion;
- The development of two methodologies based on deformable models, namely active shape models and active appearance models, that allow for characterizing the shape of the vocal tract for speech production assessment of European Portuguese language in magnetic resonance images;
- The application of the developed methods for the modeling of the vocal tract, a study of the best parameters to use in each model depending on the quality of the images as well as the qualitative and quantitative evaluation of the segmentation results by using the models referred to;
- A comparison between the performances of active shape models and active appearance models and discussion on the advantages and disadvantages of each of these models;
- The study of models using images with quality 1.5T and 3T (superior) and posterior evaluation and comparison of the segmentation results;
- Presentation of a realistic use case of application of the previous methodology that helps imaging experts and speech therapists by effectively reducing the amount of time spent on manually segmenting the vocal tract in new images;
- The study of four methodologies based on background subtraction models to perform segmentation of the human silhouette in new images;

- The development of a methodology based on active shape models for characterizing the silhouette of a human subject from an image sequence, which can be used later to perform their segmentation in new images; in addition to information on the contour of the silhouette, the developed method also integrates information on specific anatomical points such as the position of the head, shoulders, elbows, right and left hip positions, knees and feet;
- An application of the previously mentioned methods for the modeling of the human silhouette in four different datasets, a study of the best parameters to use in each model depending on the quality of the images as well as the qualitative and quantitative evaluation of the segmentation results;
- A comparison between the performances of the models and discussion on the advantages and disadvantages of using each model built.

1.4. List of Publications

In the scope of this Thesis, the following publications were produced:

1.4.1. Book Chapters

- M.J.M. Vasconcelos, J.M.R.S. Tavares. Human Motion Segmentation using Active Shape Models. Accepted in *Computational and Experimental Biomedical Sciences: Methods & Applications*, Lecture Notes in Computational Vision and Biomechanics, Springer, October 2013.
- S.M. Rua Ventura, M.J.M. Vasconcelos, D.R.S. Freitas, I.M.A.P. Ramos, J.M.R.S. Tavares. Speaker-specific articulatory assessment and measurements during Portuguese speech production based on Magnetic

Resonance Images. In *Language Acquisition*, ISBN: 978-1-61209-569-1, Nova Science Publishers, Inc., pp. 117-138, May 2012.

1.4.2. Journal Articles

- M.J.M. Vasconcelos, S.M. Rua Ventura, D.R.S. Freitas, J.M.R.S. Tavares. Inter-speaker speech variability assessment using statistical deformable models from 3.0 Tesla magnetic resonance images. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, ISSN: 0954-4119 (print) - 2041-3033 (online), Professional Engineering Publishing, DOI: 10.1177/0954411911431664, Volume 226, Issue 3, pp. 185-196, March 2012.
- M.J.M. Vasconcelos, S.M. Rua Ventura, D.R.S. Freitas, J.M.R.S. Tavares. Towards the Automatic Study of the Vocal Tract from Magnetic Resonance Images. *Journal of Voice*, ISSN: 0892-1997, Elsevier, DOI: 10.1016/j.jvoice.2010.05.002, Vol. 25, No. 6, pp. 732-742, November 2011.
- M.J.M. Vasconcelos, S.M. Rua Ventura, D.R.S. Freitas, J.M.R.S. Tavares. Using Statistical Deformable Models to Reconstruct Vocal Tract Shape from Magnetic Resonance Images. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, ISSN: 0954-4119 (print) - 2041-3033 (online), Professional Engineering Publishing, DOI: 10.1243/09544119JEIM767, Volume 224, Number 10 / 2010, pp. 1153-1163, 2010.
- J.M.R.S. Tavares, F.J.S. Carvalho, F.P.M. Oliveira, I.M.S. Reis, MJ.M. Vasconcelos, P.C.T. Gonçalves, R.R. Pinho, Z. Ma. Computer Analysis of Objects' Movement in Image Sequences: Methods and Applications. *International Journal for Computational Vision and Biomechanics*, ISSN: 0973-6778, Serials Publications, Vol. 2, No. 2, pp. 209-220, July-December 2009.

- M.J.M. Vasconcelos, J.M.R.S. Tavares. Segmentation Methods for Human Motion Analysis from Image Sequences. *ICCES*, ISSN: 1933-2815, Tech Science Press, DOI: 10.3970/icces.2009.010.003, Vol. 10, No. 1, pp. 3-4, 2009.
- M.J.M. Vasconcelos, J.M.R.S. Tavares. Methods to Automatically Build Point Distribution Models for Objects like Hand Palms and Faces Represented in Images. *Computer Modeling in Engineering & Sciences*, DOI: 10.3970/cmes.2008.036.213, Tech Science Press, ISSN: 1526-1492 (print) - 1526-1506 (online), vol. 36, no. 3, pp. 213-241, 2008.

1.4.3. Conference Papers

- S. R. Ventura, M. J. M. Vasconcelos, D. R. Freitas, I. M. Ramos, J.M.R.S. Tavares. Speech Articulation Assessment Using Dynamic Magnetic Resonance Imaging Techniques. In *VipIMAGE 2011 - III ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing*, Real Marina Hotel & Spa, Olhão, Algarve, Portugal, 12-14 October 2011, ISBN: 978-0-415-68395-1, e-ISBN: 978-0-203-85830-1, Taylor and Francis, pp. 225-231.
- M.J.M. Vasconcelos, S.R. Ventura, J.M.R.S. Tavares, D. R. Freitas. Analysis of Tongue Shape and Motion in Speech Production using Statistical Modeling. In *SEECCM 2009 - 2nd South-East European Conference on Computational Mechanics*, ISBN: 978-960-254-683-3, pp. 96-103., 22-24 June 2009, Island of Rhodes, Greece.
- M.J.M. Vasconcelos, J.M.R.S. Tavares. Métodos de Segmentação de Imagem para Análise da Marcha. In *3º Congresso Nacional de Biomecânica*, ISBN: 978-989-96100-0-2, pp. 563-564, Instituto Politécnico de Bragança, Bragança, Portugal, 11-12 Fevereiro 2009.
- M.J.M. Vasconcelos, J.M.R.S. Tavares. Human Motion Analysis: Methodologies and Applications. In *CMBBE 2008 - 8th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering*, 6 pag., Porto, Portugal, 27th February-1st March 2008.

1.4.4. Conference Abstracts

- M.J.M. Vasconcelos, J.M.R.S. Tavares. Human Motion Segmentation using Active Shape Models. In *ICCEBS2013 - International Conference on Computational and Experimental Biomedical Sciences*, 1 pag., Hotel Marina Atlântico, Ponta Delgada, S Miguel Island, Azores, October 20-22, 2013.
- M.J.M. Vasconcelos, J.M.R.S. Tavares. Segmentation Methods for Human Motion Analysis from Image Sequences. In *colloquium 511 - Biomechanics of Human Motion, New Frontiers of Multibody, Techniques for Clinical Applications*, pp.19, University of the Azores, Ponta Delgada, Azores, Portugal, March 9-12, 2011.
- M.J.M. Vasconcelos, S.M. Ventura, D.R.S. Freitas, J.M.R.S. Tavares. Modelling and Segmentation of the Vocal Track during Speech Production by using Deformable Models in Magnetic Resonance Images. In *6th World Congress on Biomechanics*, pp. 538, Singapore Suntec Convention Centre, 1-6 August 2010.
- M.J.M. Vasconcelos, S.M. Rua Ventura, D.R.S. Freitas, J.M.R.S. Tavares. Segmentation of the Vocal Tract in Magnetic Resonance Images using Deformable Models. In ICCES'10 - International Conference on Computational & Experimental Engineering and Sciences, 28 March - 1 April 2010, Las Vegas, USA.
- J.M.R.S. Tavares, M.J.M. Vasconcelos, R.R. Pinho. Motion Tracking in Images based on Stochastic Filters and Optimization. In *CMBBE2010 - 9th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering*, ISBN: 978-0-9562121-3-9, Arup, 1 pag., Westin Hotel, Valencia, Spain, 24-27 February, 2010.
- M.J.M. Vasconcelos, J.M.R.S. Tavares. Segmentation Methods for Human Motion Analysis from Image Sequences. In *ICCES'09 - International Conference on Computational & Experimental Engineering and Sciences*, ISBN-10: 0-9717880-9-X, ISBN-13: 978-0-9717880-9-1, Tech Science Press, pp. 141-142, 8-13 April 2009, Phuket, Thailand.

- M.J.M. Vasconcelos, J.M.R.S. Tavares. Methodologies for Human Detection in Image Sequences. In *3DMA-'08 - 10th Meeting of the technical group on '3D Analysis of Human Movement' of the International Society of Biomechanics*, 2 pag., Santpoort-Amsterdam, the Netherlands, October 28th - 31st, 2008.
- M.J.M. Vasconcelos, J.M.R.S. Tavares Image Segmentation for Human Motion Analysis: Methods and Applications. In *8th. World Congress on Computational Mechanics (WCCM8) / 5th. European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2008)*, ISSN: 978-84-96736-55-9, 2 pag., Venice, Italy, June 30 - July 5, 2008.

1.5. Organization of Scientific Events

During this PhD project, the following symposium was organized:

- J.M.R.S. Tavares. Y. Zhang, M.J.M Vasconcelos. Image Processing and Analysis. *Symposium within the International Conference on Computational Experimental Engineering & Sciences (ICCEES) 2009*, Phuket, Thailand, April 2009.

2

Image Analysis of the Human Vocal Tract and Silhouette

The task of finding and identifying objects in an image is trivial for humans, despite the fact that the object image can vary depending on its viewpoint or size. Even if the object is occluded or the image quality is low, humans can still easily recognize it. It is a natural task that we are prepared for from the moment we are born.

Computer vision studies how to reconstruct, interpret and understand a 3D scene from its 2D images in terms of the properties of the objects present in the scene [Schalkoff 1989]. Therefore, the ultimate goal of computer vision is to model and replicate human vision using computational algorithms at different levels. For this, it is necessary to combine the knowledge of distinct fields such as computer sciences, electrical engineering, mathematics and biology in order to understand and simulate the human vision system [Szeliski 2010].

In addition, the ability to extract points from an image that can characterize an object in an image or image sequences is of extreme importance for the computer vision field. These characteristics may involve many tasks in image analysis such as object detection, shape recognition, image registration and object understanding. Motivated by its wide range of applications, object analysis has been evolving considerably over recent decades, with various examples of

applications found in medical imaging, human gait analysis or surveillance systems [Umbaugh 2010].

Considering this background, this Thesis focuses on the development of computer algorithms for image analysis. Particular attention is given to two applications: the human vocal tract and human motion. Therefore, this chapter is dedicated to reviewing the state-of-the-art of computational algorithms for image analysis that have been used for both applications.

Regarding the subject of the vocal tract, presented throughout the first section of this chapter, the structure is as follows. The vocal tract anatomy will be reviewed, followed by a review of the imaging techniques used to obtain the picture of the complete vocal tract. Then, models that have been used to represent the shape of the vocal tract are described, followed by a summary of speech production studies that have been developed in the various languages. The section ends with some examples of the importance of vocal tract modeling.

The research available on the subject of human motion analysis will be described in the second section. The most current research is presented, followed by the methodologies used to study motion detection. Next, the techniques developed for human motion tracking are described. Current understanding of motion, along with multiple applications in human analysis is addressed next and the existing datasets are referred to. Finally, the challenges that human motion studies still have to overcome are pointed out.

2.1. Vocal Tract

Verbal communication is the most common, familiar and frequently used form of human interaction, which results from the organized and synchronized work of a set of anatomic organs. The articulation is a result of the activity of a set of organs: the vocal tract that modifies their position and shape during air expulsion (expiration), producing different sounds and consequently, distinct acoustic representations.

Since the beginning of studies of this nature, the process of speech production has attracted human interest aiming at reaching a deeper understanding and modeling of all the mechanisms involved by taking both morphological and speech acoustic aspects into consideration. The main anatomic aspects and the physiology of the vocal tract are common to all individuals. However, the mechanism engaged in human speech production is complex and unique due to the variety of anatomical structures that compose the vocal tract, implying that any computational modeling developed needs to be flexible so as to permit accurate individual characterizations [Stone 1991; Benesty et al. 2008].

Two different approaches have been used to determine the shape of the vocal tract: direct methods based on geometrical measurements of the vocal tract; and indirect methods based on acoustic inversion [Ball et al. 2008]. Among the direct measurement methods, several imaging methods have been used to obtain a complete picture of the vocal tract, like X-ray Radiography, Computed Tomography or Magnetic Resonance Imaging [Thimm et al. 1999; Ventura et al. 2009; Bakhshaee et al. 2013]. Indirect methods, in contrast, determine the vocal tract shape through acoustic data, either from a speech signal or from the acoustic response of the vocal tract. As image analysis is the basis of this Thesis, this dissertation will mainly focus on works based on direct approaches.

By assembling the former facts, it is straightforward that the study of the speech production is a multidisciplinary subject. To name a few, it involves subjects like: medicine, with the anatomic and functional study of the vocal tract organs; engineering, particularly informatics with the construction of the vocal tract models for speech and acoustics analysis; medical imaging, with the improvement and application of computational image techniques that can be used in the study of the vocal tract during speech production; phonetics, in the study of the production and perception of speech and sounds; and speech therapy, with the assessment of anatomic and physiological aspects related to communication disorders, language and speech [Beautemps et al. 1995; Baer et al. 1991; Apostol et al. 1999].

2.1.1. Anatomy

Various organs play important roles in the production of numerous speech sounds, functioning in an organized, i.e. articulated, manner in order to change the shape and length of a set of air cavities - the vocal tract. Most of these organs, named articulators, are soft-tissues that execute active movements during speech production, such as the lips, tongue and velum [Benesty et al. 2008; Ventura 2012].

Sagittal data is particularly useful in the study of the entire vocal tract anatomy [Ventura et al. 2010], demonstrating the main aspects of the shape and positions of some articulators, Figure 2.1. The tongue is a large muscular organ covered by mucous membrane, located on the floor of the mouth, which is attached by muscles to the hyoid bone, mandible, styloid processes, and pharynx. Besides its key role in mastication and swallowing, the tongue has an important function in speech production because it is the articulator with the most mobility and flexibility. Its main mass is composed of a set of muscles, which permits the elongation and constriction of the entire tongue or of specific parts allowing the articulation of sounds. The tongue's structure presents a tip, which usually rests against the incisors, and margin, body, dorsum, inferior surface and root.

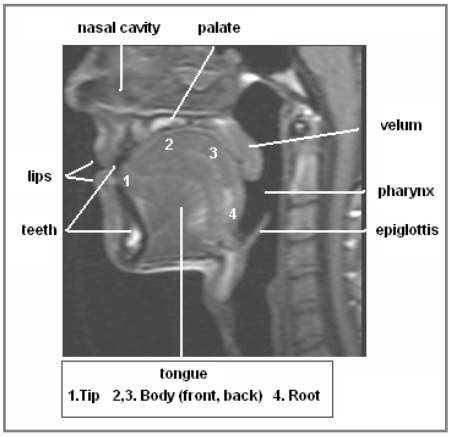


Figure 2.1 – Example of an MR sagittal slice demonstrating the vocal tract organs (from [Ventura, Freitas, et al. 2011]).

2.1.2. Imaging Techniques

Imaging techniques are methods that allow for obtaining an image of the interior of the vocal tract, and greater understanding of the positions and movements of the vocal tract organs [Stone 1991]. Two types of imaging techniques exist: structural, where the image of the structures is obtained; and tracking, where tracks are attached to important points of the vocal tract.

Examples of structural imaging techniques are X-ray Radiography, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), whereas tracking techniques include X-ray Microbeam and Electromagnetic Articulography (EMA) [Ball et al. 2008].

Radiography is a classical reference technique to study speech production. Some years after the discovery of X-rays in 1895, radiographic images of the upper vocal tract were taken by phoneticians [Harshman et al. 1977]. Since then, many techniques have been developed, from still pictures to video.

The advantages of using such an imaging technique include the possibility of obtaining the full sagittal view of vocal tract articulators during running speech with optimal temporal resolution, about fifty images per second. However, due to ethical concerns of the potential side effects of radiation exposure inflicted on the examiners, X-ray imaging technology is now rarely practiced [Xue et al. 2006]. Another limitation of this technique is related to the fact that volumetric information cannot be obtained, since X-ray can only take plain images of the speech mechanisms [Xue et al. 2006].

Nonetheless, studies have been developed using x-ray imaging, but they rely on existing images and databases previously acquired. The ATR X-ray film database is the largest X-ray database available for speech research [Munhall et al. 1995], with 25 different films of 55 minutes in total, it offers nearly 100,000 images. Other cineradiographic databases exist, such as the French database [Arnal et al. 2000], developed by the Strasbourg Institute of Phonetics and the Grenoble Institute of Speech Communication. Research by [Thimm et al. 1999;

Höwing et al. 1999; Fontecave et al. 2005; Fontecave Jallon et al. 2009] are examples of studies that combine x-ray images to comprehend speech production phenomena.

Another imaging technique used to acquire shape information about the vocal tract is Computed Tomography [Perrier et al. 1992; Inohara et al. 2010; Bakhshaee et al. 2013]. This imaging technique allows making the distinction between bones, soft tissues and air, but does not allow for discriminating different soft tissues. Moreover, the scanning speed of image acquisition is much higher than MRI. On the other hand, it has the serious disadvantage of requiring significant ionizing radiation doses, and for this reason, few studies adopt this imaging technique.

From the imaging modalities that have been used to study the vocal tract's shape and articulators, MRI has been the most commonly accepted. Its key advantages include the quality and resolution of soft-tissues and the use of non-ionizing radiation [Avila-Garcia et al. 2004; Engwall 2003]. In addition, it allows for morphologic measurements in static as well as dynamic studies [Avila-Garcia et al. 2004; Engwall 2003]. Its main drawback is that a supine position is generally required that might interfere with normal position of the vocal tract during speech [Engwall 2000a].

Due to the lengthy data acquisition time of the early MR imaging systems, the first studies were restricted to vowels and some consonants [Baer et al. 1991; Narayanan et al. 1995]. The first comprehensive body of dimensional data on the vocal tract employing MR technology was presented by [Baer et al. 1987; Baer et al. 1991]. Another study [Crary et al. 1996] describes a dynamic MRI technique that offers several promising features for studying the configuration of the vocal tract.

With cutting-edge MR improvements, a proper 3D description on the vocal tract geometry of a speaker can be reached, both in terms of good image contrast and temporal resolution. Also, with the emerging development of rapid imaging techniques, such as synchronized sampling methods [Bresch et al. 2006] or tagged

cine-MR [Parthasarathy et al. 2007; Stone et al. 2001], the acquisition of image data regarding articulatory movements became possible. Nowadays, the acquisition of three-dimensional MR image sequences has been steady [Masaki et al. 2008] and, consequently, there are enormous expectations for attaining of image data on speech production in a more efficient and repeatable manner.

An X-ray microbeam is a tracking technique that uses very small doses of X-rays to record the movement of pellets attached to the tongue. It is a technique with little risk to the subject and it has also been used to investigate both normal and disordered speech. However, the images obtained by it show only a projection by volume, turning contour extraction very difficult. This technique is now rarely used, and like radiography, recent studies are based on previously acquired databases [Fujimura et al. 1973; Dang et al. 2002].

The other tracking technique, EMA, also provides information on speech kinematics [Perkell et al. 1992; Fitzpatrick et al. 2002; J. Kim et al. 2014]. EMA has the advantage of providing the same information as the X-ray microbeam, with higher temporal resolution but without using radiation. Another advantage is that it allows for a cleaner speech audio-recording environment. The main disadvantage, as in X-ray microbeam, is that EMA requires placing small connector coils in and on the speaker's mouth, with the main problem being that relatively few subjects can tolerate this due to the sensitivity of the soft palate.

2.1.3. Vocal Tract Models

Information about the vocal tract shape and dimensions are essential to a full understanding of the articulatory and acoustical processes involved in speech production. The search for realistic and precise models to represent the vocal tract is long and several methods have been studied as reported below.

In [Harshman et al. 1977], the authors present a model to describe the tongue shapes of English vowels for five speakers. The model is based on a full x-ray measurement procedure that is reduced to a few underlying components by

means of the statistical techniques of factor and principal-component analysis. A year later, [Shirai et al. 1978] used statistical analysis of real data to describe the position of the articulatory organs. In [Maeda 1988], a factor analysis of the lateral shapes of the vocal tract is described.

The procedure to obtain the vocal tract shape [Story et al. 1996] from MR image sets included the segmentation of the airspace from the surrounding tissue, shape-based interpolation to generate the reconstruction of the airspace and analysis of the cross-sectional area. In [Kagawa et al. 1997], the authors model the vocal tract wall by an assemblage of spline functions, which is deformable around the points of interest. Also, in [Stone et al. 1997] a principal component analysis was used to examine sagittal tongue contours for five English vowels constructed from ultrasound images. In [Thimm et al. 1999] the segmentation of the vocal tract is done in an iterative manner. First, the teeth were tracked using two specialized histogram normalization techniques combined with a pattern-matching algorithm that also gives the position of the palate. Then, the normalization of the position of the vocal tract is used to track the throat and the lips. Finally, background subtraction is used to enhance the contrast of the tongue and configure its deformation. The referred segmentation procedure was optimized for X-ray images showing the side view of the vocal tract.

In 2000, a 3D tongue model was developed by [Engwall 2000b] within the *Kungliga Tekniska Hogskolan* (KTH) 3D vocal tract project using manually extracted tongue contours from MR images of a reference subject producing 43 sustained Swedish articulations. The extraction of the articulatory model's parameters was done by decomposing the geometrical points of the tongue in linear components, through a Linear Component Analysis, where the factors to be extracted were imposed on the model using MR images articulatory measures. Two years later, in [Badin et al. 2002], a database of 3D geometrical description of tongue, lips and face was established for a speaker sustaining a set of French allophones. For this, data from MRI, along with a video with and without a jaw splint were used. An important finding of this research was that, most 3D geometry of tongue, lips and face could be predicted from their midsagittal contours, at least for speech assessment purposes. Indeed, the knowledge acquired from midsagittal data and from traditional 2D models is far from obsolete. A

complement of the previous models was presented in [Serrurier et al. 2005] based on the same French subject, achieving a final articulatory model of the shape of the complete vocal and nasal tracts. For this research the 3D surface that defines each organ of the vocal tract was extracted from MRI and CT images, where the 2D contours were manually extracted from the corresponding images and later expanded to 3D. Then, principal component analysis was applied to the set of organ surfaces to uncover the two main uncorrelated articulatory degrees of freedom for the velum's movement.

A region growing technique was also explored to model the vocal tract shape [Behrends et al. 2003]. Here the authors first matched teeth phantoms to the MRI dataset to perform the segmentation and reduce human time expense. The segmentation method places a seed inside the vocal tract that expands until it reaches its walls. The expansion is based on gray-level comparison between the mean gray-level value of the segmented region and the neighborhood pixels of its contour until a defined difference value is achieved. The vocal tract midline is also computed by using a modified 1D-Kohonen algorithm to calculate the characteristic area functions. Later in [Carbone et al. 2008], the same technique was used to segment 100 2D vocal tract contours over a European Portuguese Database achieving Pratt Indices from 84% to 100%. A recent study [Silva et al. 2013] also uses region growing to segment the vocal tract, this time on real-time MR image sequences.

In [Mollaei et al. 2008], the authors use radiography images to obtain the vocal tract shape and calculate the median line through center of gravity of the contours of the vocal tract. To obtain the vocal tract shape model, the authors followed the approach of [Maeda 1988; Beautemps et al. 1995] and presented a model based on simple polynomials.

A semi automatic technique for facilitating the extraction of vocal tract contours is described in [Fontecave Jallon et al. 2009]. The method combined the manually acquired geometrical data for a small number of key images and used a similarity measure based on the low-frequency Discrete Cosine Transform components of the images to automatically index the other images. Finally, the acquired contours are combined to reconstruct the movements of the entire vocal tract. In [Bresch et al. 2008] an unsupervised regional segmentation technique was

adopted to track the contours of about a dozen tract variables, as the lip and velum aperture or the tongue tip constriction.

More recently, CT images were used to create models of the vocal tract in [Bakhshaee et al. 2013]. The intensity contrast between air and tissue is high in CT images, so the vocal tract boundaries were clearly identified. A semi automatic segmentation procedure was used based on a region-based sectioning method: the threshold values were determined after an indication of the user of the region of interest.

Statistical models were also used to represent the vocal tract shape. In [Avila-Garcia et al. 2004] active shape models and Hough transform were employed to extract the shape of dynamic MR images. In contrast, shape deformation techniques to define and extract the vocal tract in static MR images are presented in [Vasconcelos et al. 2010; Vasconcelos et al. 2011; Vasconcelos et al. 2012]. Active Shape Models (ASM) [Cootes et al. 1995] and Active Appearance Models (AAM) [Cootes et al. 1998] are used to define the vocal tract shape model. Details on these models are given in the next chapter of this Thesis. Last year, in [Raeesy et al. 2013] a method of automatic landmark tagging by recursive boundary subdivision was applied to obtain the corresponding sets of landmarks on the vocal tract contours. Here, an active-orientated shape model technique was adopted to recognize and delineate the shape of the vocal tract in standardized MR images. To avoid the task of manually positioning the landmarks, a recursive boundary subdivision approach [Rueda et al. 2011] was used.

2.1.4. Studied Languages

The study of the vocal tract has been used for speech assessment in many different languages, namely English [Masaki et al. 1996], Swedish [Engwall et al. 2000], French [Soquet et al. 1996], Japanese [Takemoto et al. 2004], German, European Brazilian [Gregio 2006; Pontes et al. 2009] and European Portuguese (EP) [Martins et al. 2008; Ventura et al. 2009].

For the English, [Harshman et al. 1977] describes the tongue shape of ten English vowels for five subjects. An inventory of speaker-specific, three dimensional, vocal tract air space shapes was done by [Story et al. 1996], corresponding to a particular set of vowels and consonants, namely 12 vowels, 3 nasal and 3 plosives.

Regarding the Swedish language, in [Engwall et al. 2000], a tongue model was developed for MR images of a reference subject producing 43 artificially sustained Swedish articulations.

In [Badin et al. 2002], the geometry of vocal organs is measured on one subject uttering a corpus of sustained articulations in French. Later, a more complete study used a corpus of 46 French phonemes [Serrurier et al. 2005]. Also focusing on French, in [Clément et al. 2007], MR images of the vocal tract were obtained from one subject during sustained production of three French point vowels with short scanning duration. The manually traced boundaries of the vocal tract served to obtain estimates of the area functions, which were later used as input for a speech simulation system.

Also, in [Takemoto et al. 2004], five sets of volume data were acquired to extract the vocal tract shape during sustained production of the Japanese vowels from one subject. Two MRI corpora of one male subject were acquired for German [Birkholz et al. 2006]: the first from sustained phonemes of 18 sagittal slices; and the second from dynamic sequences of three utterances.

The first study concerning European Portuguese (EP) language dates back to 1997 [Teixeira et al. 1997], in which a software tool is presented to study Portuguese vowels. Since then, other studies from the same group have been presented [Teixeira et al. 2001; Teixeira et al. 2005; Martins et al. 2008]. Meanwhile, and since research on EP remained scarce, other studies have been presented [Ventura et al. 2008; Ventura et al. 2009; Vasconcelos et al. 2010; Ventura, Freitas, et al. 2011; Vasconcelos et al. 2012; Ventura 2012].

2.1.5. Applications

Static MRI measurements have shown to be representative of dynamic speech and demonstrated that the articulations in the MRI data are hyperarticulated [Engwall 2000a]. The hyperarticulation in artificial sustained articulations is a natural consequence of the subject aiming to produce as clear examples as possible of each articulation, thus enlarging the important distinctions and reducing coarticulatory effects at the tongue contour.

In [Xue et al. 2006], the vocal tract dimensions of White American, African American and Chinese male and female speakers were compared. A total of 120 adult subjects were studied and six dimensional parameters of the speakers' vocal tract cavities were measured with acoustic reflection technology.

Another interesting result obtained from the comparison of the vocal tract area functions from the same speaker in 1994 and 2002 [Story et al. 1996; Story 2008] showed that the data were not identical, suggesting a different vocal tract setting for producing the same sounds in distinct time lines. Differences were observed in the cross-sectional area variation along the vocal tract axis as well as differences in the vocal tract length. The obtained data showed that the vocal tract shape may be highly variable for the same target vowel depending on the particular setting used by the speaker, which is very useful for understanding intra-speaker variability.

Recently, [Laukkanen et al. 2012] investigated the effects of using a straw in voice training and therapy. The results indicated that, in fact, exercising with a straw helps establish a speaker's formant cluster, which increases loudness and improves vocal economy. Also, the results obtained from CT and acoustic studies in [Guzman et al. 2013] suggested that vocal exercises with increased vocal tract impedance lead to increased vocal efficiency and economy.

The potential of using active models is highlighted in [Miller et al. 2014], by improving the knowledge and understanding of factors underlying structural and functional variations of vocal tract structures. With this, better treatments and therapies can be developed for those with speech difficulties as well as more effective strategies for improving vocal technique in professional singers.

2.2. Human Motion

Human motion analysis usually follows a general framework: human detection, human tracking and human understanding [L. Wang, Hu, et al. 2003], as depicted in Figure 2.2. The first step involves the extraction of low-level features, which aims at segmenting and identifying regions corresponding to people or body parts from the remaining portion of the image. Only afterwards, through an intermediate-level, can the tracking of such objects be done. The final step of human motion analysis consists of understanding the behavior of the former features along the image sequence, where activity recognition is performed, such as gesture, action or interaction recognitions.

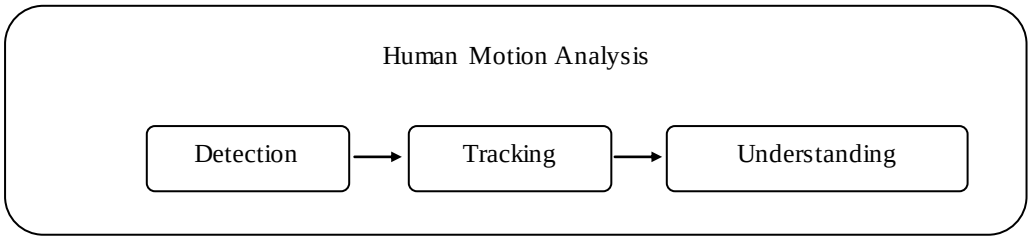


Figure 2.2 – Human motion analysis framework.

2.2.1. Surveys

Human motion analysis has been an active topic in computer vision over the years and the series of survey papers in the literature confirm this. Thus, this section is dedicated to several surveys that have been done regarding this field.

The first significant review on human motion analysis was probably due to [Aggarwal et al. 1994], who reported on the developments on non-rigid motion analysis examining the trends in the research of articulated and elastic motion. In both trends, motion recovery methods that use no *a priori* shape models are separated from those that use model based approaches. A year later, [Cedras et al. 1995] presented the developments in the computer vision aspect of motion-based recognition, starting with the extraction of motion information and its

organization into models and continuing to the problem of matching unknown input with a model. After presenting overall methods for motion, the authors focused on methods for the tracking and recognition of human motion. Here, volumetric and stick figure models are described, as well as 2D and 3D tracking methods, and finishing the survey with human motion recognition methodologies.

Years later, [Aggarwal et al. 1999] presented another overview of the tasks involved in human motion analysis, covering the work prior to 1998. The work focused on three major areas related to interpreting human motion such as motion analysis involving human body parts, tracking of human motion using single or multiple cameras, and recognizing human activities from image sequences. In the same year, [Gavrila 1999] published a survey on visual analysis of gestures and whole-body movement, where both involved articulated objects. The work was organized according to the dimensionality of the tracking space, 2D or 3D, and the type of models used, with or without explicit shape models.

Later, [Moeslund et al. 2001] presented a survey on computer vision based on human motion capture from 1980 into the first half of 2000. The focus was on a general overview based on the four primary functionalities of motion capture processing, including initialization, tracking, pose estimation and recognition. Throughout the paper, a number of general assumptions used in this field were identified and suggestions for future research were offered. Considering the substantial progress towards human motion tracking and reconstruction, the same authors presented a sequel of their former work, based on more than three hundred papers, while maintaining the same functional taxonomy [Moeslund et al. 2006].

A survey of the various studies related to the human tracking and body parts was presented by [J. J. Wang et al. 2003] in addition to approaches related to modeling behavior using motion analysis. In the same year, [L. Wang, Hu, et al. 2003] presented another review of the subject giving special emphasis to human detection, tracking and activity understanding. The authors also discussed some research challenges and future directions. Later in [Poppe 2007] the author summarized the characteristics of markerless vision-based human motion analysis, dividing the analysis into a modeling and an estimation phase.

More recently, [Aggarwal et al. 2011] presented another review, this time concentrating on high-level activity recognition methodologies designed for the analysis of human actions, interactions and group activities. Finally, the most recent survey comes from [Nissi Paul et al. 2014], now focused on human walking motion. It presents approaches used for human detection, various tracking methods, with different approaches for pose estimation and pose analysis. It also includes the use of unsupervised systems to understand walking motion.

In [Jaimes et al. 2007], a broader view of the state-of-the-art of multimodal human-computer interaction is given. Motivated by the multidisciplinary nature of this field, the authors discussed major approaches and issues from a computer vision perspective, discussing topics from large-scale body movement, gesture recognition and gaze detection, to facial expression or emotion analysis. On the contrary, a survey specifically directed to human motion tracking for rehabilitation is given in [Zhou et al. 2008]. The work reviewed the development of human tracking systems and their application in stroke rehabilitation.

Recently, Human Behavior Analysis and Understanding (HBA/HBU) has been increasingly of interest to computer vision researchers, [Chaaraoui et al. 2012] dealing with state-of-the-art HBA/HBU from an Ambient Intelligence perspective, focusing especially on indoor scenarios and techniques designed for Ambient-Assisted Living purposes. [Metaxas et al. 2013] focused on reviewing research in the area of human Nonverbal Communication Computing, and, particularly, motion analysis developed to address this problem.

So far the surveys presented only consider images acquired from a conventional camera; however, recently, depth sensors have made a new type of data available and researchers have been on top of it. In [Chen et al. 2013], a review with the advantages of this type of imaging is presented as well as the main published research for analyzing human activity.

2.2.2. Motion Detection

The process of extracting from a sequence of images the regions corresponding to humans is commonly called segmentation. This process is usually based on either temporal or spatial information [L. Wang, Hu, et al. 2003; Moeslund et al. 2006]; however, efforts have been made to use both types of information in order to obtain enhanced results.

Temporal Information

In most cases, temporal information is used when background and camera are static, thus it is possible to obtain the movement of the subjects in analysis through the differences between images from a sequence. The simplest method of motion segmentation consists of subtracting the current image from the previous one.

One of the most typical methods for motion segmentation is background subtraction, which consists of subtracting the intensity or gradient of each pixel of an image sequences from a reference background image. This methodology usually presents good results in controlled environments but is quite sensitive in outdoor environments, illumination changes and presence of shadows. Therefore, in order to overcome these limitations several techniques have been presented ever since [Piccardi 2004].

Another type of methodology for motion segmentation based on temporal differencing uses a pixel-wise difference between two or three consecutive frames in an image sequence to extract moving regions. For instance, [Zhao et al. 2006] presented a method that extracts contours of moving objects mainly by combining gradient information with three-frame-differencing and connectivity-testing-based noise reduction. The former method has the advantage of being very adaptable to environments that are dynamic and has low computational complexity.

Optical flow is also another interesting alternative method to detect moving objects and consists of detecting of interesting points, features or blobs based on pixel values and further linkage using flow vectors. In [Min et al. 2004], the authors presented a method for extraction and temporal segmentation of multiple

motion trajectories in human motion. They begin by detecting candidate motion locations in every frame and then obtain the motion trajectories by combining the significant motion points with the color optical flow based tracker results. Optical flow methods have the advantage that they can be used in the presence of camera motion.

Spatial Information

Motion detection using spatial information can be done either through threshold or statistical approaches.

Thresholding relies on a simple process based on special environmental assumptions. For instance, it can be used when: the subject wears dark clothes against a different background; an infra-red camera is used, like in [Goubet et al. 2006]; or the subject wears special markers on key points.

Statistical approaches, instead, use appearance assumptions together with subtraction methods. As an example of a statistical approach, Wren et al. [Wren et al. 1996] first proposed a running Gaussian average method. The idea of this method is to fit a Gaussian probability density function on the last N pixel values, updating independently each pixel by running a cumulative average. Another technique reported by [Stauffer et al. 1999; KaewTraKulPong et al. 2002] consists of using a mixture of Gaussians to model the background with a shadow detection scheme incorporated into it. This statistical method seems to learn faster and more accurately and adapts well to changing environments.

Other authors, like [Elgammal et al. 2003], explored the use of a non-parametric model based on kernel density estimation with the fast Gauss transform to model the background distribution. The disadvantage of this method is related to the high associated computational cost.

In [Oliver et al. 2000] the authors adaptively built an eigenspace to model the background, where the eigenspace model describes the range of appearances observed in the image sequences, such as the lighting or weather variations throughout the day.

Spatio-Temporal Information

Some researchers have been developing methods for human detection that use both spatial and temporal information to obtain better results [Amer 2003; Ahmed et al. 2006].

In [Rui et al. 2000], the authors presented an algorithm that takes input video sequences, computes frame to frame optical flow, projects the flow fields into a basis set using singular value decomposition analysis and detects temporal discontinuities in the trajectories of the basis coefficients over time.

In [L. Liu et al. 2005], the authors proposed a video segmentation method that integrates two major components: short-based video segmentation and object-based segmentation. The key-frame extraction is used to provide a compact video representation that contains the salient and video content objects, and then a joint spatiotemporal video segmentation is used to extract the objects through a generative clustering method.

The authors in [Dimitrijevic et al. 2006] presented a template-based approach to detect human body poses, in which the templates consist of short sequences of 2D silhouettes obtained from motion capture data. The method combines silhouette matching with motion information and statistical relevance. The technique presented good results in both indoor and outdoor sequences though they were acquired with a moving camera.

2.2.3. Motion Tracking

Tracking unconstrained movement of a human in image sequences is extremely challenging [Ning et al. 2004]. It is a difficult but important task in human motion analysis.

Most of the methodologies used for human motion analysis are model-based, for example, shape models like stick figures, 2D contours or volumetric models [Aggarwal et al. 1999]. Other examples of methodologies include active contour-based and feature-based, which will be described in this section.

Model-based Tracking

The most simple and useful model that represents the human body structure is the stick figure model, which connects sticks through joints. In [Guo et al. 1994], a model with ten sticks articulated with six joints is constructed, using silhouettes as image features. It classifies the stick figure motion into walking, running and other motions through a neural network. In this work, only one moving object, the person, exists in the scene and also a stationary background and parallel projection are assumed. A independent tracking view of the human figure is achieved in [Karaulova et al. 2002], where a stick figure representation is used to model the human body, and Hidden Markov Models are used to encode the model dynamics.

In [Mikić et al. 2003], the authors presented an integrated system for automatic acquisition of the human body model and motion tracking using input data acquired from multiple synchronized video streams. The system performs the tracking on the 3D voxel reconstructions computed from the 2D foreground silhouettes. The human body model used consists of ellipsoids and cylinders and is described using a twisted framework resulting in a non-redundant set of model parameters.

In [L. Wang, Tan, et al. 2003], the authors described a method for automatic person recognition from body silhouette and gait, which combines a background subtraction procedure with a simple correspondence method to segment and track spatial silhouettes of a walking figure. In order to reduce the computational cost during training and recognition, simple feature selection and parametric eigenspace representation are used.

A different possibility is to use a motion model to accomplish human tracking. For example, in [Ning et al. 2004], a motion model was constructed from the semi-automatically acquired training data and motion constraints were explored by analyzing the dependency of joints. Both of them were later integrated into a dynamic model to reduce the size of the sample set.

[Cheung et al. 2005] constructed a body model from scratch using simple joint connection knowledge of the body without using any *a priori* shape model. The skeletal structure is registered using video sequences of the person moving

their limbs and shape information is extracted from the body parts directly from the silhouette and color images. The tracking algorithm works very well for relatively simple motions, but for complex motions it suffers from the problem of local minima.

A markerless tracking algorithm of human motion from multiple camera views was proposed in [Kehl et al. 2006], their solution integrates features such as edges, color and volumetric reconstruction capable of correctly categorizing self and partial occlusions. A stochastic optimization is later used to find the best match between the articulated body model and the computed features.

Another type of methodology consists of using the appearance to construct the human model. In [Ramanan et al. 2007] an automatic system to track the articulations of persons from a video sequence is presented. It starts by constructing a model of appearance of each person in a video and then tracks it by detecting this model in each frame. It describes two approaches that learn their appearance: the first is a bottom-up algorithm that groups together candidate body parts found throughout a sequence and the second is a top-down approach that constructs appearance models from convenient poses. The system can count distinct individuals, is capable of identifying and tracking different people, and is able to recover when it loses their track as well. Results are shown in frames of unscripted indoor and outdoor activity, a feature-length film and legacy sports footage.

The authors in [Rius et al. 2008] used a stick figure model which learns the 3D variability of human posture using a set of training sequences. They developed a matching algorithm based on Dynamic Programming to establish mapping between postures from different motion cycles. Then, the model is trained, a mean walking performance is automatically learnt and the statistics about the observed variability of the postures and motion direction are also computed. As an alternative, in [Meeds et al. 2008], a probabilistic stick-figure model is presented that uses a nonparametric Bayesian distribution over trees for its prior structure.

Also, 2D contours are often used to detect humans in image sequences; for example, in [Korč et al. 2008] a three-step algorithm was presented, which detects human candidates, validates the model of a human and finally tracks the model in

consequent frames. The model adopted is a six-link model with an articulated head that can cope with a frontal view of a person. It starts using simple means to find a human candidate within a region of interest and afterward validates it using an extended biped human model.

In [Freifeld et al. 2010], a contour person model is defined, that captures natural shape and pose variations. The deformations from a training template are used to describe changes in shape due to camera view, body shape and articulated pose. In this study, only frontal bodies were used in the 2D model and the inclusion of other views require an inference method to search the discrete set of views.

Another type of algorithm is the articulated Iterative Closest Point (ICP) such as the one presented by [Corazza et al. 2010]. An articulated subject-specific model was created from direct measurement of the subject outer surface using either a laser scan or visual hull frame. The tracking approach employed a minimization scheme over an ICP algorithm with six degrees of freedom for each body joint.

In [Straka et al. 2011], the authors use silhouette images to construct a volumetric model of the human body and extract a skeletal graph from it. Then, by using a matching algorithm based on geodesic distances, they assign labels to the end-nodes of the graph and later determine the inner-nodes. At the moment, they are working on handling cases in which the skeletal graph becomes corrupt as a result of the arms being too close to the upper body.

Recently, [Yoo et al. 2011] explains a markerless system to describe, analyze and classify human gait motion. The authors use a sequential set of 2D stick figures to represent the motion. Features based on motion parameters are determined and measured in order to characterize the gait patterns. This research began back in 2002 [Yoo et al. 2002], when the authors explored the possibility of extracting the gait signature and kinematic features guided by known anatomy.

Active-Contour-based Tracking

Active contours use boundary detectors that iteratively move towards the final solution according to the combination of image and optional user-guidance forces [Terzopoulos et al. 1988; Blake et al. 1998; Szeliski 2010]. These models consider the object boundary as a single, connected structure with underlying geometric representations.

An example of active contours, snakes, which were first introduced by [Kass et al. 1988], represent a salient image feature as a parametric curve that can move under the influence of internal forces and aims to minimize the energy associated with the curve. The main drawbacks of these models were the failure to detect nonconvex objects and its sensitivity to initialization.

An alternative model for edge detection, derived from the classical active contour model, is the geodesic active contour model, introduced by [Caselles et al. 1997]. These models are derived from geometric functional models and are non-linear, leading to inefficient implementations. For instance, explicit Euler schemes for the geodesic active contour limit the numerical step for stability. The former drawback was overcome in [Goldenberg et al. 2001] and [Paragios et al. 2000], who also improved the model by using level sets to describe contours and a gradient descent algorithm to optimize it.

In [Kwon et al. 2007], the authors combine geodesic active contour models with a mean-shift algorithm. The initial curve in each frame is re-localized near the human region and resized enough to include the target object, to reduce the number of iterations and handle large object motion.

An active model which characterizes regional and structural features of a target object such as shape, texture and color is presented in [Jang et al. 2000]. The model is capable of adapting itself dynamically to an image sequence in order to track a non-rigid moving object.

Level set techniques were presented in [McInemey et al. 1999; McInerney et al. 2000]. They are a development of the conventional snakes in the sense that they enable topological flexibility among other features. While many methods rely on edges, this method [Chan et al. 2001] optimally fits a two-phase piecewise

constant model to the given image, where the boundary is represented with a level set function, which can handle topological changes more easily than explicit snake models.

In [Xin et al. 2004], a contour tracking algorithm for video captured using mobile cameras of different modalities is proposed. The algorithm used Bayesian inference based on the probability density functions of texture and color features. In addition, it adopted the features of both object and background regions in the level set evolution model. The limitation of this model is that pixel values are treated as if they were independent for posterior probability estimation, making the contour sensitive to disturbances caused by similarities of color or texture between the object and the background.

In [Cremers 2006], the authors develop dynamic statistical shape models for implicitly represented shapes, capable of capturing the temporal correlations which characterize deforming shapes such as the consecutive silhouettes of a walking person. A Bayesian formulation for level set based image sequence segmentation imposes the statistically learned dynamic model as a shape prior to segmentation processes.

More recently, [Hu et al. 2013] also presented a framework for active contour-based visual tracking using level sets. The framework includes: contour-based tracking initialization for the first frame; a color-based contour evolution algorithm to achieve tight and smooth contours; adaptive shape-based contour evolution to make the shape model flexible; dynamic shape-based contour evolution to obtain more accurate contours; and abrupt motion handling, by incorporating particle swarm optimization into level set evolution. The proposed method can be used to track object contours, regardless of whether the camera is stationary or moving, and it can deal effectively with videos with abrupt motions.

Active shape models can be also applied to the tracking of non-rigid objects, as human models, in a video sequence. These models are a compact form for which the shape variety and the color distribution of an object class can be both taught in a training phase [Cootes et al. 1995]. Its compactness results from principal component analysis and *a priori* shape information from the training set. In [Koschan et al. 2003], a hierarchical realization of an enhanced active shape

model for color video tracking is presented and performances of hierarchical and non-hierarchical implementations are studied. Active shape models are also used by [D. Kim et al. 2006], this time for a panoramic image obtained from multiple sensors. In [Dou et al. 2007], an ASM-based people tracking system is implemented in a reconfigurable hardware to accelerate the ASM algorithm, since it requires great computational power for real time people tracking.

In [Rathi et al. 2005], the authors formulate a particle filtering algorithm in the geometric active contour framework that can be used for tracking moving and deforming objects. Occlusion is dealt with by incorporating shape information into the weights of the particles. Experiments of a walking couple sequence are shown.

Feature-based Tracking

[Comaniciu et al. 2000] present an approach based on visual features such as color and texture, whose statistical distributions characterize the object of interest. Mean shift iterations are then employed to find the target candidate most similar to a given object model.

In [Gonzalez et al. 2003], the authors presented a robust feature-based tracking method of human motion. The approach presented enables tracking motions of different body parts without articulated body models and their initialization by using a standard point-wise tracker modified for robustness and grouping image points undergoing the same rigid motions.

An approach that combines prior knowledge regarding a person's motion with human body kinematics constraints was presented in [Sappa et al. 2005]. The approach computes feature point trajectories and uses the peaks and valleys of these to detect key frames, where both legs are in contact with the floor, and those key frames allow the association of the motion models with each joint. The authors also presented experimental results considering different video sequences.

In the work of [Ekinci et al. 2005], the authors begin modeling the background to obtain the silhouettes of the moving objects and identify persons according to some shape features, like the bounding box ratio or the second moment of the silhouette. Then, the person's center of mass is calculated as well as the local maxima of the filtered signal from the distances between the center and the silhouette. In the tracking process, correspondences are established by using the minimum cost criteria. In addition, this work also presents results for motion classification, namely normal walking and running.

The problem of probabilistic modeling of human motion is addressed in [Rogez et al. 2006], by combining several 2D views. A multi-view Gaussian Mixture Model (GMM) is fitted to a feature space made of shapes and stick figures manually labeled. The temporal and spatial constraints are considered to construct a probabilistic transition matrix, which is used to limit the feature space only to the most probable models from the GMM.

In [Tanaka et al. 2007], the authors extracted the skeleton from the captured volume data using the thinning process and then converted it into an attributed graph using an exemplar based-approach. Body parts are identified from each curved line in the skeleton through a graph-matching algorithm.

[Sundaresan et al. 2008] propose a method of articulating objects tracking in the Laplacian Eigenspace. It is shown that Laplacian Eigenmap transform is suitable for extracting the 1D object and for segmenting the different chains in the joints and then k-dimensional splines are used to model these smooth 1D curves in the eigenspace. After segmentation has been performed, the skeleton is estimated using the registration of the nodes along the 1D curve.

In [Nascimento et al. 2008], the authors represent the human body by its center of mass and a bank of switched dynamic models is used to describe the trajectory of the pedestrian in the image sequence. The models are trained offline from hand segmented video sequences in a supervised way.

In [G. Liu et al. 2010], the authors present a computer vision system capable of automatically tracking the movements of skaters on a large-scale complex and dynamic rink. The authors chose to use Scale Invariant Feature Transformation (SIFT) features, due to its invariance to viewpoint changes, large geometric

transformation and changes in illumination. The tracking system incorporated the hierarchical model based on contextual knowledge into the unscented Kalman filter.

[Saini et al. 2012] presented vision based human motion tracking using a non-linear dimension reduction charting technique. The human body structure was extracted using a Gaussian mixture model based silhouette descriptor and joint configuration in manifold space belonging to low-dimensional space. The mapping between the two spaces was done with a relevance vector machine. The main goal of the descriptor is to reflect a silhouette as a set of intelligible regions in the 2D space like foreground pixel locations.

In [Barbu 2014], the authors determine what moving image objects represent pedestrians by testing several conditions related to human bodies by detecting the skin regions from the movie frames. A Histogram of Orientated Gradients (HOG) based template matching process was used in the tracking stage. While most methods use HOG in the detecting stage of human motion, the work referred to uses it in the tracking stage and found that it works better than other template matching approach.

2.2.4. Motion Understanding

The improvement of the interaction between men and machines is essential for the growth of human motion analysis. A wide variety of disciplines, from surveillance to medicine, have been interested in this subject as described next.

For instance, in surveillance systems, human motion analysis can be used to identify suspicious movements of persons in a parking lot or to monitor the actions of individuals and classify their nature in a commercial space. These types of activity can require a considerable effort from human operators, since it is common to have several cameras in a parking lot or a shopping area that must be analyzed simultaneously.

In [Nascimento et al. 2005], the authors proposed an algorithm to model, segment and classify human activities in a constrained environment by using switched dynamic models. In [Cucchiara et al. 2005], the authors analyze human behaviors by classifying the posture of the monitored person and consequently detecting corresponding events and emergency situations, like a fall. The former approach can be applied to monitor people at home, especially the elderly with limited autonomy, and define potential emergency situations.

In sports, a biomechanical analysis of movements of athletes can help them understand and improve their performances or even facilitate the recovery process after injuries.

In [Krosshaug et al. 2007], the authors present a model-based image matching technique to extract kinematic characteristics of three typical anterior cruciate ligament (ACL) injury situations, which can provide valuable information on the mechanisms for ACL injuries in sports. Another example, is the Football Interaction and Process Model system (FIPM), which can acquire action models, infer action-selection criteria and determine player and team strengths and weaknesses [Beetz et al. 2005].

Another application area where human motion analysis plays an important role is Gait Analysis. Gait can be defined by motor behavior consisting of integrated movements of the human body. The cyclical pattern of corporal movements can be linked to a specific individual, allowing human recognition through it.

In [Begg et al. 2005], the authors show results that support vector machines are able to automatically recognize gait patterns of elderly and young people. Both histogram and Poincaré plot diagram features are effective in discriminating the two age groups, which can indicate that such plots might be useful in detecting movement abnormalities or for monitoring improvements in walking performances because of treatment or intervention in a clinical procedure.

In [Rius et al. 2008], the authors propose an action specific model which automatically learns the variability of 3D human postures observed in a set of

training sequences. Dynamic Programming techniques are used to synchronize the training sequences and, as a result, they obtain an action model with a representative manifold for the action; namely, the mean performance, the standard deviation from the mean performance and the mean observed direction vectors from each motion subsequence of a given length. The resulting model can be used for gait recognition applications such as in the identification of a subject when performing an action by observing only a reduced motion portion of it.

A gait recognition system for human identification is proposed by [Rani et al. 2010], using a modified independent component analysis (MICA). Background modeling is done in order to segment the moving objects and then a skeleton operator is used to track the moving silhouettes of a walking figure. The sequence of silhouette images is used to train the MICA based on eigenspace transformation and the gait features are recognized based on a self-similarity measure. The work of [Arantes et al. 2011] presents a framework that merges four different models of human movement, using a fusion model to improve classification. Each model was based on specific image segmentation of the human silhouette and extracted global information on tri-dimensional, bi-dimensional, boundary and skeleton motion. The results suggest that the framework is capable of recognizing people by their gait.

In medicine, the study of human motion can also be extremely valuable. In [Davis III et al. 1991] the authors described a clinical gait analysis system used at the Newington Children's Hospital, and also presented the clinical testing protocol and the algorithms used. Over ten years later, in [Šimšík et al. 2005], motion analysis was used in the study of spondylolisthesis. [Schubert et al. 2005] also carried out a motion study in patients with Parkinson's disease. In [Goulermas et al. 2005], the authors present tests of an extensive range of dimensionality reduction and robust classification techniques for linking pathological plantar hyperkeratosis and functional biomechanical foot data.

Another area of application of human motion analysis is Computer Graphics. In [Remondino et al. 2004], the authors present a framework for the modelling and animation of human characters from monocular videos. [Nguyen

et al. 2005] describes a real-time system for capturing humans in 3D and placing them into a mixed reality environment, where the images of the subject are constructed using a robust and a fast shape-from-silhouette algorithm.

2.2.5. Motion Datasets

In the last 20 years, several datasets dedicated to human motion have been created to serve as input data for various research problems. In this section, a brief review on available datasets for video-based human activity and action recognition is presented.

A summary of the most relevant datasets used to date for human analysis is provided in Table 2.1 and explained next. A human motion analysis dataset should gather three conditions to be complete, these being: 1) to have sufficiently high-resolution images to capture details; 2) a high frame rate to detect movements; 3) multiple cameras to see the subject from varying viewpoints [Chaquet et al. 2013]. For more detailed information about the subject it is also important to have a motion capture (MoCap) system, to capture their movement.

The first dataset on human data was created in 1998 by Visual Computing Group, University of California, San Diego (UCSD) [Little et al. 1998]. It is an outdoor sequence where the subjects walk parallel to a homogenous wall and perpendicular to the camera. Later, a more complete set, with more subjects and more walking directions, appears with the Institute of Automation Chinese Academy of Sciences (CASIA)-A dataset [L. Wang, Tan, et al. 2003], the former National Laboratory of Pattern Recognition (NLPR) gait database. Here, the subjects walk on a straight-line path at free cadences and in three different viewing angles with respect to the camera.

The Carnegie Mellon University Motion of Body database (CMU MoBo) [Gross et al. 2001], focuses on biometric identification of humans from their individual characteristics. The database contains four different styles of walking (slow, fast, inclined and carrying a ball) performed on a treadmill by 25 subjects.

Table 2.1 – Summary of most relevant datasets for video-based human analysis.

Dataset Name	Year	Video #(cams,fps)		Resolution	# (Subjects, Actions,Sequences)			Ground truth
UCSD	1998	1	30	640x480	6	1	42	Outdoor
CASIA-A	2001	1	25	352x240	20	1	240	Outdoor
CMU-MoBo	2001	6	30	640x480	25	1	100	-
CMU-Mocap	2001	1	15	320x240	144	109	2605	Mocap 12cams, 41mpp
NADA	2004	1	25	160x120	25	6	2391	-
CASIA-B	2005	11	25	320x240	124	1	13640	Indoor
HumanID	2005	2	30	720x480	122	1	1870	Outdoor
HDM05	2005	1	25		5	>70	1500	MoCap 12cams,~40mpp
HumanEva	2009	7	60	640x480	4	6	56	MoCap 12cams,195mpp
MuHAVi	2009	8	25	720x576	7	17	119	Manual annotation
MPI08	2010	8	40	1004x1004	4	14	54	3D laser scan 5 sensors
UMPM	2011	4	50	644x484	30	>15	36	MoCap 14cams, 37mpp

The CMU motion capture database [CMU Graphics Lab 2001] is the most extensive dataset of publicly available motion capture data. The only drawback of this database is related to the lack of calibration information, required to project the 3D models into the images, rendering it unsuitable for evaluating video-based tracking performance.

In 2004, the department NADA from the Computer Science and Communication at Stockholm University introduced the NADA database by [Schuldt et al. 2004], containing 6 types of different actions (walking, jogging, running, boxing, hand waving and hand clapping) in four different scenarios, three outdoors and one indoor with homogeneous backgrounds. The drawback of this database is the image resolution of only 160x120 pixels.

A massive multiview gait database was created in 2005, called CASIA-B [Yu et al. 2006]. With 124 subjects, it is the most complete database mentioned here, with data captured from 11 views and two variations, namely clothing and carrying positions. The data is captured indoors and only for walking action. Besides the video files, the database also provides human silhouettes extracted from the videos. A similar dataset is the HumanID [Sarkar et al. 2005]. The difference is that data is acquired outdoors and only two video cameras are used. Also in the same year, motion capture data was recorded at the Hochschule der

Medien (HDM05) [Muller et al. 2007], being a very well structured dataset, with the same action being performed many times.

The HumanEva dataset [Sigal et al. 2010] provides ground-truth data to assist in the evaluation of algorithms for estimating pose and tracking human motion. The database contains six different actions performed by four subjects wearing natural clothes. The motivation for wearing natural clothes, instead of using the typically tight-fitting motion capture suits, is to obtain natural images, containing the complexity posed by moving clothing. However, a drawback is that natural clothes provide ground truth motion caption data that is less accurate compared with data collected by traditional methods.

The Multicamera Human Action Video Dataset (MuHAVi) dataset [Singh et al. 2010] provides multi-camera human action video data with manually annotated silhouette data. The advantage of this dataset is that the data has been collected in a site with challenging lighting conditions provided by multiple sources of night streetlights [Singh et al. 2010].

The Indoor Motion Capture (MPI08) dataset [Pons-Moll et al. 2010], constructed by the University of Hannover, contains a wide variety of human motion. The database is recorded in an indoor setup and consists of 4 subjects performing 14 different motion patterns, such as walking, jumping or throwing. As a complementary data source to visual information, 5 inertial sensors were fixed to the body extremities to obtain accurate limb orientations.

The Utrecht Multi-Person Motion (UMPM) benchmark [Van der Aa et al. 2011] includes synchronized motion capture data and video sequences from multiple viewpoints for multi-person motion including multi-person interaction. This dataset has also the advantage of including static objects in the scene like a table or a chair, in order to allow testing methodologies regarding occlusion cases. Another differentiator aspect of this dataset is that the video cameras do not face each other directly, to prevent similar silhouettes. In addition to this, supplementary data such as background images and the assignment of 3D MoCap data to a specific subject are provided.

For a more complete survey on video-based human action and recognition, [Chaquet et al. 2013] is the most appropriate reading suggestion where a total of

68 datasets are referred to, with 28 belonging to heterogeneous and 40 to specific human actions. Comparison and classification of such datasets are also provided.

2.2.6. Challenges

Image segmentation methods related to human motion must deal with several challenges [Nissi Paul et al. 2014] such as:

- dynamic backgrounds: for instance, when the camera is moving; illumination conditions that can vary throughout the image sequences;
- visibility problems: when the subject does not remain inside the workspace or is partially occluded by other elements of the scene;
- image sequences with more than one subject in the workspace at the same time.

The development of methods that can deal with all these problems simultaneously is not a straightforward task so it is common to make some assumptions. However, each day increasingly robust and accurate methods are being developed.

If human segmentation in video sequences is challenging, the tracking task is no different. A few of these challenges that are worth mentioning are:

- the complex non-rigid structure of the human body, with its high number of degrees of freedom [Ning et al. 2004]. It has many joints and each body part can move in a wide range around its corresponding joint;
- dealing with frequent self-occlusion of body parts due to ambiguity inherent in 3D to 2D projection [L. Wang, Tan, et al. 2003], which will provide valuable information about hidden motion;
- usage of markers for motion capture are only suitable for well-controlled environments [Kolahi et al. 2007; Sandau et al. 2014];
- shape and appearance variation of the human movements due to clothes;
- abrupt motion handling;
- tracking response in practical time [Nikolaidis et al. 2009].

3

Vocal Tract Active Models: Application to the European Portuguese Language

The application of statistical models, such as deformable and active models, to characterize and reconstruct the vocal tract during speech production was taken into consideration in the present Thesis. The development of active models to represent the vocal structures from a global perspective is here presented.

Actually, the studies regarding the analysis of the vocal tract during the production of European Portuguese sounds are still scarce. Hence, the motivation to extend the knowledge of this particular language from the representation of the vocal tract trough active models during speech production was also here explored.

In this chapter, active models were built to segment the shape of the vocal tract in new images for speech production assessment of the European Portuguese language. The first section provides an explanation about the sounds of European Portuguese language. The second section briefly describes Point Distribution Models, Active Shape Models and Active Appearance Models. In section three, one describes the image datasets used as well as the Magnetic Resonance Imaging protocols. The fourth section presents the implementation and the various models constructed for the study of the shape of the vocal tract and appearance during speech production and an example of their application to real studies. Finally, in section five, the results are discussed and conclusions are presented.

This chapter contains work and results of the article "Analysis of Tongue Shape and Motion in Speech Production using Statistical Modelling" by Maria João M. Vasconcelos, Sandra R. Ventura, João Manuel R. S. Tavares and Diamantino R. Freitas published in June 2009 in the Proceedings of 2nd South-East European Conference on Computational Mechanics. It also contains work and results of the journal paper "Using Statistical Deformable Models to Reconstruct Vocal Tract Shape from Magnetic Resonance Images" by Maria João M. Vasconcelos, Sandra M.R. Ventura, Diamantino R. S. Freitas and João Manuel R.S. Tavares published in 2010 in the Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine. It also contains work and results of the journal paper "Towards the Automatic Study of the Vocal Tract from Magnetic Resonance Images" by Maria João M. Vasconcelos, Sandra M. R. Ventura, Diamantino R. S. Freitas and João Manuel R.S. Tavares published in November 2011 in Journal of Voice. At last, it contains work and results of the journal paper "Inter-speaker speech variability assessment using statistical deformable models from 3.0 Tesla magnetic resonance images" by Maria João M. Vasconcelos, Sandra M. R. Ventura, Diamantino R. S. Freitas and João Manuel R. S. Tavares published in March 2012 in the Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine.

3.1. European Portuguese Language

According to the International Phonetic Alphabet (IPA), the European Portuguese (EP) language consists of a total of 30 sounds: nine vowels, two diphthongs and nineteen consonants[International Phonetic Association 1999]; in addition, EP is one of the most widely spoken languages.

In the sound productions of both EP vowels and diphthongs, the articulators remain sufficiently spaced out allowing air flow to pass freely and almost without obstacles. The main difference between the oral vowels configuration comes from the position of the lips and tongue. Vowels are classified in four different classes: open, close, mid and central, according to the position of the tongue, lip's

projection and the mouth aperture [International Phonetic Association 1999; Ventura 2012]. The tonic system of Portuguese European is composed by nine oral vowels:

- Open front unrounded vowel [a]: from the Portuguese word /c*a*sa/ (home);
- Mid central unrounded vowel [ɐ]: from the Portuguese word /c*a*da/ (each);
- Open-mid front unrounded vowel [ɛ]: from the Portuguese word /p*é*/ (foot);
- Close-mid front unrounded vowel [e]: from the Portuguese word /m*e*do/ (scare);
- Open-mid back rounded vowel [ɔ]: from the Portuguese word /p*ó*/ (dust);
- Close-mid back rounded vowel [o]: from the Portuguese word /f*o*rça/ (strength);
- Close front unrounded vowel [i]: from the Portuguese word /r*i*so/ (laughter);
- Close back rounded vowel [u]: from the Portuguese word /t*u*/ (you);
- Mid central unrounded vowel [ɨ]: from the Portuguese word /sed*e*/ (thirst).

With regards to the production of the EP vowels [i, e], the tongue moves to higher frontal positions, and in the case of the EP vowels [o, u], the tongue moves to more elevated backward positions. The EP sound [a] is produced when the tongue is to be found in a central and mid-low position.

A diphthong is formed by one vowel that is pronounced stronger (the vowel itself) and one that is pronounced weaker (identified semivowel) [International Phonetic Association 1999; Ventura 2012]. The sounds [a, e, o] regularly work as vowels, and the sounds [i] and [u] regularly work as semivowels.

Phonetically, the EP vowels and diphthongs are regarded as being long and somewhat continuous sounds, classified from the front to the back of the mouth and from the higher to the lower tongue positions.

There are two main classes of consonants: plosives and fricatives [International Phonetic Association 1999; Ventura 2012]. Plosive consonants consist on sounds in which air stream from the lungs are interrupted by a complete closure in some part of the vocal tract. The occlusion may be done with the tongue (blade [t], [d]), or body [k], [g]), lips ([p], [b]), or glottis ([ʔ]). Plosives consonants contrast with nasals, where the vocal tract is blocked but airflow continues through the nose, as in /m/ and /n/. In fricatives sounds, on the contrary, the air usually passes through a narrow constriction that causes the air to flow turbulently and thus create a noisy sound.

The other classes of consonants that are found in the majority of languages, namely nasals, "liquids" and vowel-like approximants, are voiced in the overwhelming majority of cases.

Consonants can be classified according along three major dimensions: (1) place of articulation, (2) manner of articulation and (3) voicing [International Phonetic Association 1999; Ventura, Freitas, et al. 2011; Ventura 2012]. One of the major differences among consonants is in the accompanying action of the larynx, with the most larynx settings that allow air to flow freely between the vocal folds versus one in which the vocal folds vibrate to produce regular voicing.

In this manner, it is relatively easy to identify the distinctive features of the sounds produced. As far as the EP fricative consonants are concerned, the places of articulation are:

- Voiceless labiodental [f]: from the Portuguese word /*f*é/ (faith)
- Voiced labiodental [v]: from the Portuguese word /*v*ê/ (see)
- Voiceless alveolar [s]: from the Portuguese word /*s*ol/ (sun)
- Voiced alveolar [z]: from the Portuguese word /ca*s*a/ (home)
- Voiceless post-alveolar [ʃ]: from the Portuguese word /*j*á/ (already)
- Voiced post-alveolar [ʒ]: from the Portuguese word /*ch*ave/ (key)

Nasality, by opposite, is a complex feature, defined by the lowering of the velum to open the velopharyngeal port, which induces strong and complex changes in the vocal tract acoustical behavior. The EP Language is especially rich in nasal sounds – both vowels and consonants.

3.2. Point Distribution Model

Point Distribution Models (PDM) have been widely used in the statistical modeling of objects to analyze its shape configurations from a set of training images [Cootes, Taylor, et al. 1992]. Thus, it describes the mean shape of the object modeled together with admissible variations in relation to the same mean shape.

In the building process of a PDM, each shape of the object to be modeled, presented in the training set, should be represented by a set of labeled landmark points. These points should reflect important aspects of the object's boundaries or interior. In order to study the variation of the coordinates of the landmark points of the training shapes it is necessary that they are aligned [Cootes, Taylor, et al. 1992]. An example of an alignment method to be used is given in [Oliveira et al. 2008], based on dynamic programming.

Hence, given the co-ordinates (x_{ij}, y_{ij}) of each landmark point j of the shape i of the modeled object, the shape vector is

$$\boldsymbol{x_i} = (\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{in}}, \boldsymbol{y_{i1}}, \boldsymbol{y_{i2}}, \dots, \boldsymbol{y_{in}})^{\boldsymbol{T}}, \qquad 3.1$$

where $i = 1, \dots, N$, with N representing the number of shapes in the training set and n the number of landmark points. Once the shapes are aligned, the mean shape and the variability can be found. The modes of variation characterize the manners in which the landmarks of the shape tend to move together and can be obtained by applying Principal Component Analysis (PCA) to the deviations from the mean. Thus, it is possible to rewrite each shape vector x_i as

$$x_i = \bar{x} + P_s b_s, \qquad 3.2$$

where x represents the n landmark points of the new shape of the modeled object, (x_k, y_k) is the position of landmark point k, $\bar{x}$ is the mean position of landmark points, $P_s = (p_{s1} \quad p_{s2} \quad \cdots \quad p_{st})$ is the matrix of the first t modes of variation, p_{si} corresponds to the most significant eigenvectors in a PCA of the position variables, and $b_s = (b_{s1} \quad b_{s2} \quad \cdots \quad b_{st})^T$ is a vector of weights for each variation mode

of the shape. Each eigenvector describes the manner in which linearly correlated x_i move together over the training set. Equation 3.2 represents the Point Distribution Model of an object and can be used to generate new shapes of it. Further details about the construction of PDM can be found in references [Cootes, Taylor, et al. 1992; Cootes and Taylor 1992].

The local gray level environment of each landmark point can also be considered in the modeling of an object [Cootes and Taylor 1992]. Thus, statistical information is obtained about the mean and covariance of the gray level values of the pixels around each landmark point. This information is used in the PDMs variations: to evaluate the match between landmark points in Active Shape Models (ASM) and to construct the appearance models in Active Appearance Models (AAM), as it will be explained next.

3.2.1. Active Shape Model

The combination of PDM and the gray level profiles for each landmark of an object can be used to segment this object in new images through the Active Shape Models, which is an iterative technique for fitting flexible models to objects represented in images [Cootes and Taylor 1992].

The referred technique is an iterative optimization scheme for PDMs allowing initial estimates of pose, scale and shape of an object to be refined in a new image. The used approach is summarized on the following steps: 1) at each landmark point of the model calculate the necessary movement to displace that point to a better position; 2) calculate changes in the overall position, orientation and scale of the model which best satisfy the displacements; 3) finally, through calculating the required adjustments to the shape parameters, residual differences are used to deform the shape of the model [Cootes et al. 1995].

In [Cootes et al. 1994] the authors presented an improved active shape model using multiresolution. So, the proposed method first constructs a multiresolution pyramid of the input images by applying a Gaussian mask, and afterwards studies the gray level profiles on the various levels of the pyramid built, making active models faster and more reliable.

3.2.2. Active Appearance Model

This approach was first proposed in [Cootes et al. 1998] and allows for the building of texture and appearance models. These models are generated by combining a model of shape variation (a geometric model), with a model of the appearance variations in a shape-normalized frame. The statistical model of the shape used it is also described by Equation 3.2.

To build a statistical model of the gray level appearance, each example image is deformed so that its landmark points match the mean shape of the object, by using a triangulation algorithm. Then the gray level information, g_{im}, is sampled from the shape-normalized image over the region covered by the mean shape. In order to minimize the effect of global light variation, this vector is normalized in order to obtain $_g$. After applying a Principal Component Analysis to this data, a linear model called the texture model is obtained:

$$g = \bar{g} + P_g b_g , \qquad 3.3$$

where $\bar{g}$ is the mean normalized gray level vector, P_g is a set of orthogonal modes of gray level variation and b_g is a set of gray level model parameters. Therefore, the shape and appearance of any example of the modeled object can be defined by vectors b_s and b_g.

Since there may exist correlation between the shape and gray levels variations, a further Principal Component Analysis is applied to the data. Thus, for each training example the concatenated vector is generated:

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} = \begin{pmatrix} W_s P_s^T (x - \bar{x}) \\ P_g^T (g - \bar{g}) \end{pmatrix} , \qquad 3.4$$

where W_s is a diagonal matrix of weights for each shape parameter, allowing the adequate balance between the shape and the gray models. Then, a Principal Component Analysis is applied on these vectors, giving a further model:

$$b = Q c , \qquad 3.5$$

where Q is the eigenvectors of b and c is the vector of appearance parameters controlling both shape and gray levels of the model. In this way, an example object can be obtained for a given c by generating the shape-free gray level object, from the vector g, and be deformed using the landmark points described by x.

3.3. Image Datasets

For the analysis of the vocal tract configurations during sustained articulations of EP speech sounds, three datasets from two different MR acquisition systems were used, 1.5T and 3T [Ventura 2012]. Next, the description of each dataset and the corresponding MRI protocols are presented.

According to the safety procedures for MR, a questionnaire was performed for screening patients before any procedure [Ventura, Freitas, et al. 2011; Ventura 2012]. In addition, patients were previously informed and instructed about the study to be performed and informed consents were obtained.

3.3.1. 1.5T Dataset

Image acquisition was performed using a Siemens Magneton Symphony 1.5 Tesla (1.5T) system and a head array coil, with the subject lying in the supine position [Ventura et al. 2012; Ventura 2012]. Due to this experimental setup, the T1-weighted sagittal slices of 5 mm thickness were obtained by using Turbo Spin Echo Sequences, with the acquisition duration of approximately 10 s. The decrease of the slice thickness entails a low signal noise ratio, making the posterior segmentation process more complex. Subsequently, this protocol has resulted from a compromise between the signal noise ratio, the number of slices acquired and the time needed for subjects to sustain articulation successfully during image acquisition process. The acquisition parameters adopted were: field

of view equal to 150 mm; image matrix of 128x128 pixels and image resolution equal to 0.853 px/mm.

The speech corpus of the 1.5T Dataset consisted of a set of 25 MR images collected during sustained articulations of 25 EP speech sounds; that is, one sagittal image was acquired per each sound considered. The images were acquired from one young male subject in a similar manner to that which has been formerly used by other studies that use MRI to analyze the vocal tract [Serrurier et al. 2005; Badin et al. 2006; Story 2008]. The training of the subject was performed to ensure the proper production of the intended EP speech sounds and to reduce speech subject variability. Moreover, the subject in question had a vast knowledge of EP speech therapy. Additionally, the images were provided in JPEG format, with 256x256 pixels.

3.3.2. 3.0T Sounds Dataset

The image data was acquired using a Magneton Trio 3.0 Tesla (3.0T) MR system and two integrated coils (a 32-channel head coil and a 4-channel neck matrix coil), with the subjects in supine position [Vasconcelos et al. 2012; Ventura 2012]. A T1-weighted midsagittal slice of 3 mm thickness was acquired using turbo spin echo 2D sequence, and adopting the following parameters: a repetition time of 400 ms, an echo time of 10 ms, an echo train length of 5, a square field of view of 240 mm, a matrix size of 512x512 pixels, a resolution of 2.133 px/mm and a 0.469x0.469 pixel size.

The speech corpus of the 3T Dataset consisted in 25 sounds of European Portuguese language, including oral and nasal vowels, and consonants. Images were acquired from two young volunteers, one male and one female, that were trained before the MR exam to ensure the proper production of the intended sounds [Vasconcelos et al. 2012; Ventura 2012]. In order to reduce intra-speaker variability and to ensure consistency of results, 3 measurements (i.e. 3 slices per sound) were performed during the sustained sound with an overall acquisition time of approximately 8.07 seconds, resulting in 75 images for each subject. Additionally, the images were provided in JPEG format, with 512x512 pixels.

3.3.3. 3.0T Sequences Dataset

For the 3.0T Sequences Dataset, the speech corpus consisted on two sequences of sounds of European Portuguese language, in two different articulatory contexts:

- Vowel-vowel articulation;
- Set of consonant-vowel articulation during a word utterance.

The image data was acquired using a Magneton Trio 3.0 Tesla MR system and two integrated coils (a 32-channel head coil and a 4-channel neck matrix coil), with the subjects in supine position [Ventura, Vasconcelos, et al. 2011]. A Flash Gradient-Echo Sequence was used to acquire 100 midsagittal WT1 slices during 48 seconds for each repeated utterance, the followed parameters were adopted: 6 mm slice thickness, a repetition time of 6.4 ms, an echo time of 2.44 ms, a field of view of 178x220 mm, a matrix size of 156x192 pixels, a resolution of 0.873 px/mm and a 1.146x1.146 mm pixel size.

The first articulatory context included the five oral vowels [a ɛ i ɔ u] and the second, the utterance Portuguese word /pato/ (duck), with the IPA phonetic transcription [patu]. Images were acquired from two young female volunteers, without articulatory disorders, that were trained before the MRI exam to ensure the proper production of the intended sounds. The 3.0T sequence Dataset is constituted by a total of 400 images, 100 images per each sequence and from each subject [Ventura, Vasconcelos, et al. 2011]. Additionally, the images were provided in TIFF format, with 156x192 pixels.

Examples of the MR images from the datasets acquired are depicted in Figure 3.1. From these images, one may observe different vocal tract configurations for EP vowels and consonant production, as well as for some oral and nasal sounds. Comparing the several vocal tract configurations of the subjects during the articulation of the EP sounds, individual differences of vertical length and of organs morphology are revealed, although the main configurations are similar.

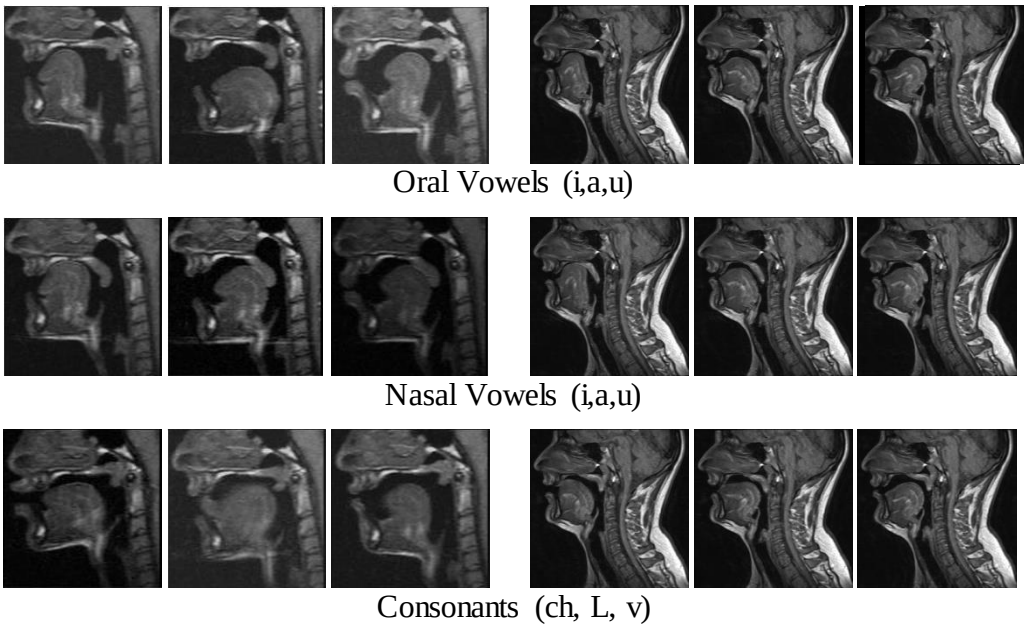


Figure 3.1 – Examples of images from the 1.5T (left) and 3.0T (right) datasets.

3.4. Models

In the present section one describes the implementation and the various models built for the study of the shape of the vocal tract and appearance during speech production as well as an example of their application to real studies.

3.4.1. Implementation

The algorithms to create the statistical deformable models were developed in MATLAB (http://www.mathworks.com), namely PDMs and ASMs, which integrates the *Active Shape Models software* [Hamarneh 1999] as basis. Additionally, in the case of the appearance models, the *Modelling and Search Software* [Cootes 2004] was used, which was built in C++ with VXL computer vision libraries (http://vxl.sourceforge.net).

An implementation for segmentation quality assessment using the Active Shape Models and Active Appearance Models built was also developed in MATLAB. In the referred implementation, the values of mean and standard deviation of the Euclidean distances between the landmark points of the final shape of the models and the desired segmentation shapes were calculated as well as the minimum and maximum values.

3.4.2. Tongue Shape Model

The first analysis focused on the tongue shape configuration during the articulation of the nine oral vowels. Thus, a statistical model, PDM, on MR images was constructed to extract the main characteristics of the tongue shape configuration using the 1.5T dataset.

A PDM was built from a set of nine images manually annotated with sixteen points along the tongue boundary, as depicted in Figure 3.2:

- Two points in the lingual *frenulum* (anterior and posterior);
- One point in the tongue's tip;
- One point in the tongue's root;
- Seven points along tongue's body;
- Five points along the inferior surface of the tongue.

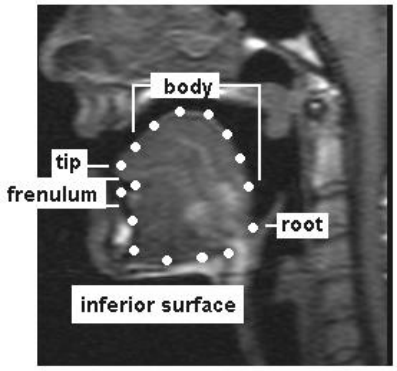


Figure 3.2 – Landmark points considered to build the tongue shape model.

From Table 3.1 one can observe that the first three modes of the shape model built could explain 90% of all shape variance of the tongue. The first five modes explain 95% of all shape variance and with only seven modes of variation it is possible to explain 99% of all shape variance of the tongue.

The effects of varying the first four modes of variation are visible in Figure 3.3. From the observation of this figure, one can depict that the first mode is associated to movements of the whole tongue along the vertical to horizontal direction. In the second mode of variation, it is possible to observe the rise of the inferior surface and of the tongue body towards the palate. The third mode of

variation translates the lowering of the tongue's tip and the advance of the tongue body simultaneously. The fourth mode of variation translates the rise and backward of the tongue dorsum. The fifth mode is related with the vertical rise of the tongue body towards the palate. The sixth mode translates the backward of the tongue's tip and finally the seventh mode is related with the diagonal movement of the whole tongue from high to lower positions.

Table 3.1 – First seven modes of variation of the model obtained and their retained percentages.

Mode of variation	Retained Percentage	Cumulative Retained Percentage
λ_1	56.453%	56.453%
λ_2	23.362%	79.815%
λ_3	10.623%	90.438%
λ_4	3.331%	93.769%
λ_5	2.454%	96.223%
λ_6	1.787%	98.010%
λ_7	1.378%	99.388%

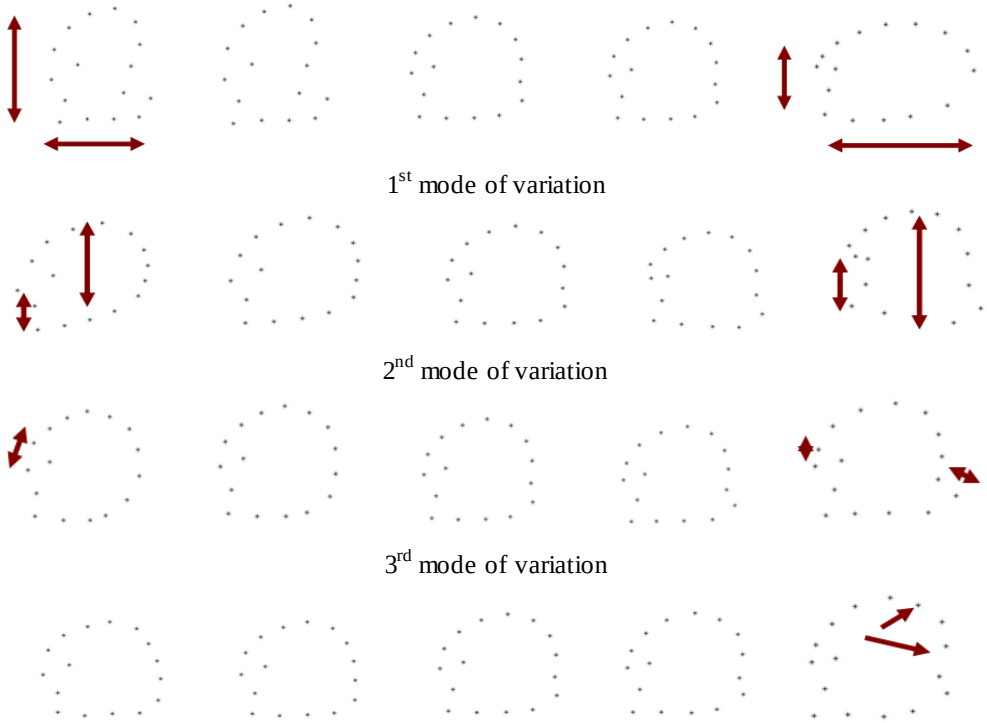


Figure 3.3 – Effects of varying each of the first four modes of variation of the tongue model (mean ± 2 standard deviation).

The PDM for the tongue model only for the EP oral vowels allowed a better understanding of the dynamic speech events involved during sustained articulations. It can be useful for speech rehabilitation purposes, namely, to recognize the compensatory movements of the articulators during speech production.

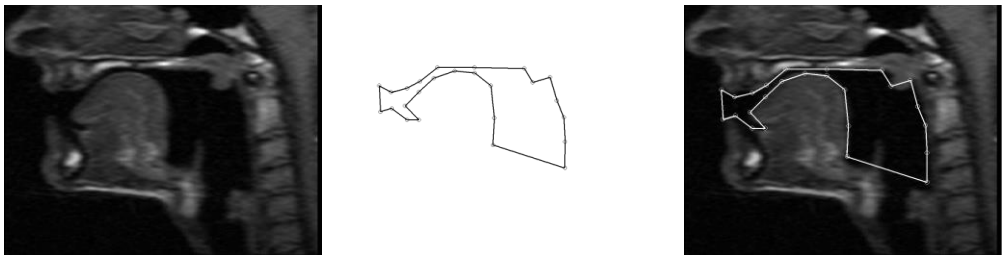
After the obtained results, it was important to explore this research towards a more complete analysis, considering the whole vocal tract anatomy and using more images, as explained in next subsection.

3.4.3. Vocal Tract Model and Sounds Reconstruction

A vocal tract model was built trough a PDM using all the 25 images from the 1.5T dataset and considering 25 manually extracted anatomical points from the vocal tract articulators, see Figure 3.4. Images were annotated by a medical imaging specialist and further cross-checked by the author, to detect possible inconsistencies or missed landmarks.

The labelling process adopted the following landmark points:

- Four points in the lips (front and back of the lip margins);
- Three points corresponding to the lingual *frenulum* and tongue's tip;
- Seven points equally spaced along the surface of the tongue;
- Seven points along the surface of the hard palate (roof of the oral cavity) placed in symmetry with the tongue points;
- One point at the velum (or soft palate);
- Three points equally spaced at the posterior margin of the oropharynx (behind the oral cavity).



a) b) c)

Figure 3.4 – a) Training image, b) landmark points selected, c) image labeled with the overlapped landmark points selected.

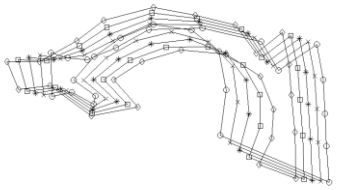
Through the analysis of Table 3.2, one may notice that the initial 7 modes of the statistical deformable model built are capable of explaining 90% of all variance of the shape of the vocal tract, while the first ten modes illustrate 95% of all variance, and to explain 99% of all variance it is necessary to use 16 modes of variation. This indicates that the PDM that has been built is capable of considerably reducing the data set that is required to represent all shapes that the vocal tract held in the images training set.

Table 3.2 – First 16 modes of variation of the model built and their retained percentages.

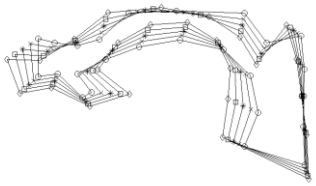
Mode of variation	Retained Percentage	Cumulative Retained Percentage
λ_1	43.598%	43.598%
λ_2	12.340%	55.938%
λ_3	10.988%	66.926%
λ_4	9.345%	76.271%
λ_5	6.947%	83.218%
λ_6	4.724%	87.942%
λ_7	2.675%	90.617%
λ_8	1.973%	92.590%
λ_9	1.428%	94.018%
λ_{10}	1.312%	95.330%
λ_{11}	0.978%	96.308%
λ_{12}	0.797%	97.105%
λ_{13}	0.654%	97.759%
λ_{14}	0.537%	98.296%
λ_{15}	0.441%	98.737%
λ_{16}	0.334%	99.071%

The effects of varying the initial four modes of variation are depicted in Figure 3.5. The first mode is associated with movements from the high front to the lower back of the tongue in the oral cavity. With regards to the second mode of variation, it is possible to observe the vertical movement of the body of the tongue towards the palate. On the other hand, the variations of the third mode have been noticed to be related with the lip movements. Finally, the fourth mode of variation reflects the approximation of the tip of the tongue to the upper alveolar region.

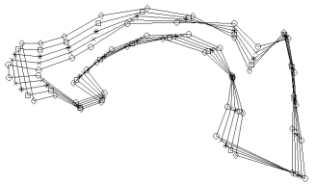
Following the construction of the vocal tract model, some sounds were chosen to be reconstructed by using the statistical deformable model built, namely five EP oral vowel sounds and the EP fricative consonants, in order to infer about the quality of the reconstruction. The main goal of the present study in this second phase was to conclude whether the modes of variation of the statistical deformable model built could be combined in order to successfully reconstruct, that is, reproduce, an EP speech sound.



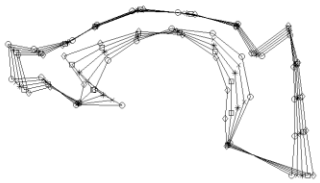
1st mode of variation



2nd mode of variation



3rd mode of variation



4th mode of variation

Figure 3.5 – Effects produced by the variation of each of the first four modes of variation of the vocal tract model built (mean ± 2 standard deviation).

In Figure 3.6, the resultant reconstructions of the shape of vocal tract related to the EP consonants [s] and [z] and the vowels [u] and [i] are depicted. In order to assess the quality of the reconstruction of the shape of the vocal tract in the articulation of EP speech sounds, the minimum, maximum and mean errors and the standard deviation of the Euclidean distances between the landmark points of the original shape and reconstructed ones must be calculated. Table 3.3 indicates such values.

Figure 3.6 – Reconstruction of the EP speech sounds [s], [z], [u] and [i]: a) original shape, b) reconstructed shape and c) both shapes overlapped.

Table 3.3 – Errors obtained of the reconstructed shapes.

	Minimum error [pixels]	Maximum error [pixels]	Mean error and standard deviation [pixels]
Consonant [s]	1.72	16.22	7.22 ± 4.13
Consonant [z]	1.04	18.25	7.11 ± 5.19
Vowel [u]	2.51	13.97	7.57 ± 3.58
Vowel [i]	1.91	18.64	9.10 ± 4.20

The sounds that revealed to be the easiest to reconstruct were the vowels [i] and [o] and the consonant [j], as it only required the combination of two variation

modes of the model built. Thus, in order to obtain the shape of the vocal tract when articulating the vowel [i], it was necessary to merge the 1st and the 4th modes. On the other hand, the reconstruction of the vowel [o] meant that the 1st and the 3rd modes needed to be united. Finally, the combination of the 1st and 8th modes enabled the reconstruction of the EP sound of the consonant [j].

Through the union of three variation modes of the statistical deformable model built, it was possible to reconstruct the shapes of the vocal tract when articulating the EP consonants [ch] and [f]. Thus, the combination of the 1st, 3rd and 7th modes permitted the reconstruction of the shape of the vocal tract associated with the consonant [ch], and by using the 1st, 2nd and 3rd modes, the shape of the vocal tract related to the consonant [f] was reproduced.

In order to obtain the shape of the vocal tract when articulating the EP consonant [v], it was necessary to bring together the 1st, 2nd, 3rd and 5th modes of variation of the statistical deformable model built whereas the consonant [s], implied the union of the 1st, 2nd, 3rd and 8th modes. On the other hand, the reconstruction of the EP vowel [a], required the combination of the 1st, 3rd, 5th, 7th and 8th modes. In the case of the EP vowel [e], the articulation of 1st, 3rd, 5th, 8th and 9th modes was adopted. Finally, the reconstruction of the EP consonant [z], implied the union of the 1st, 2nd, 4th, 6th and 8th modes, Figure 3.6.

Contrary to all expectations, the EP vowel [u] required the combination of the highest number of variation modes to reconstruct the related shape of the vocal tract. Before initiating a reconstruction study, the ones held the belief that the EP vowels were the easiest sounds to be (re)produced since that the air flows without any obstruction on the vocal tract. However, this was proven not to be the case. In fact, in order to reconstruct the sound of the EP vowel [u], the combination of the first ten modes of variation of the statistical deformable model built was required. This indicates that, from a morphological and dynamic point of view, the EP vowel [u] is not as simple to reconstruct as one would initially believe.

In terms of phonation, fricative consonants are classified as either being voiceless or voiced, implying that the sounds are produced with or without the vibration of the vocal cord. The process of reconstruction used throughout this

work has also proven that the dynamics associated with the production of these sounds are distinct: on the one hand, the articulatory points are located in diverse positions; on the other hand, the combination of the variation modes has been proven to be a more complex phenomenon. This may be exemplified by the fact that the reconstruction of the voiceless consonant [s] implied the combination of 4 variation modes (1^{st}, 2^{nd}, 3^{rd} and 8^{th}) whereas the voiced sound [z] required five variation modes (1^{st}, 2^{nd}, 4^{th}, 6^{th} and 8^{th}).

3.4.4. Vocal Tract Active Models on 1.5T MR Images

The suitability of active models to segment the shape of the vocal tract in new images is explored in the present section.

From the 1.5T dataset, 21 images were considered in the building of the statistical models, ASM and AAM, of the vocal tract's shape by using one MR image per each sound. Additionally, the other 4 MR images, related with other 4 EP speech sounds, were later used to evaluate the quality of the segmentation obtained by the Active Models built. The anatomical points considered were the ones from previous section, namely 25 manually extracted anatomical points from the vocal tract articulators, annotated by a medical imaging specialist and further cross-checked by the author, Figure 3.4 [Vasconcelos et al. 2011].

For the sensibility analysis of the Active Shape Models in terms of the percentage of the retained variance and on the dimensions of the profile adopted for the gray levels [Vasconcelos et al. 2010], ASMs were built with 95% and 99% of retained variance and with profiles for the gray levels of 7, 11 and 19 pixels. In the same way, Active Appearance Models were built with 95% and 99% of retained variance and considering 50000 and 10000 pixels for the texture model. As stopping criterion of the segmentation process, a maximum of 5 iterations on each resolution level was considered. As 4 resolution levels were defined based on the dimensions of the images, this criterion means that from the moment that the segmentation process starts to its end a maximum of 20 iterations can occur [Vasconcelos et al. 2010]. This maximum number of iterations was chosen because with the images considered it leads to excellent segmentation results. Additionally, it was verified that a lower value was not always sufficient to obtain

satisfactory segmentations and a higher value constantly leads to the same segmentation results.

In Table 3.4, the first 15 modes of variation of the active shape model built and their retained percentages are indicated. From the values presented one may conclude that the initial 7 modes are capable of explaining 90% of all variance of the vocal tract's shape under study. Additionally, one may conclude that the first 10 modes represent 95% of all variance and that the first 15 modes provide an explanation for 99% of all variance. The former results indicate that the ASM built is able to considerably reduce the data required to represent all shapes that the vocal tract assumes in the training images set.

Table 3.4 – First 15 modes of variation of the model built for the vocal tract's shape and their retained percentages.

Mode of variation	Retained Percentage	Cumulative Retained Percentage
λ_1	45.349%	45.349%
λ_2	13.563%	58.912%
λ_3	9.672%	68.584%
λ_4	9.123%	77.707%
λ_5	6.716%	84.423%
λ_6	4.674%	89.097%
λ_7	2.262%	91.359%
λ_8	1.872%	93.231%
λ_9	1.442%	94.673%
λ_{10}	1.367%	96.040%
λ_{11}	0.979%	97.019%
λ_{12}	0.701%	97.720%
λ_{13}	0.507%	98.227%
λ_{14}	0.494%	98.721%
λ_{15}	0.396%	99.227%

The effects of varying the first 6 modes of variation are depicted in Figure 3.7. This figure allows one to become aware that the first mode is associated with the movements of the tongue from the high front to the back positions at the oral cavity. With regards to the second mode of variation, it is possible to observe the vertical movement of the body of the tongue towards the palate. On one hand, the variations of the third mode are related with the opening of the lips and tongue's

movement to backward. On the other hand, the fourth mode of variation reflects the tongue's tip movement from the central position of the tongue to the alveolar ridge of the palate. Additionally, the fifth mode of variation translates the opening of the lips and the overall lateral enlargement of the vocal tract. Finally, the sixth mode is related with the movement of the tongue's body from back to front and down positions.

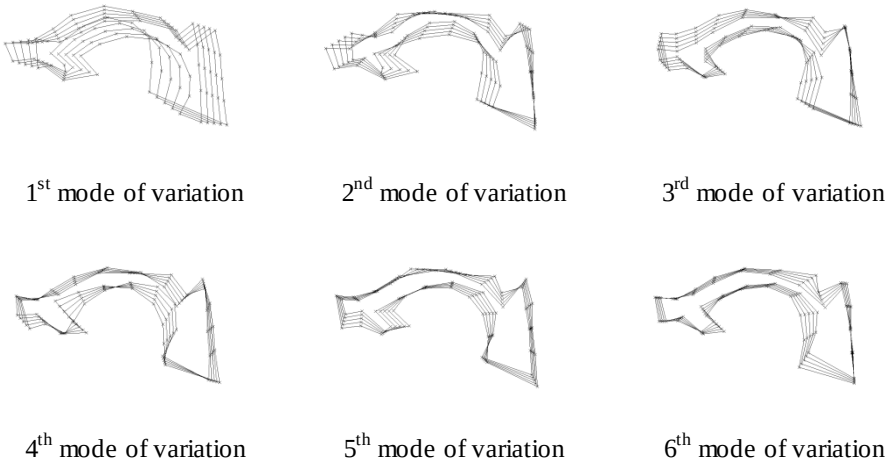
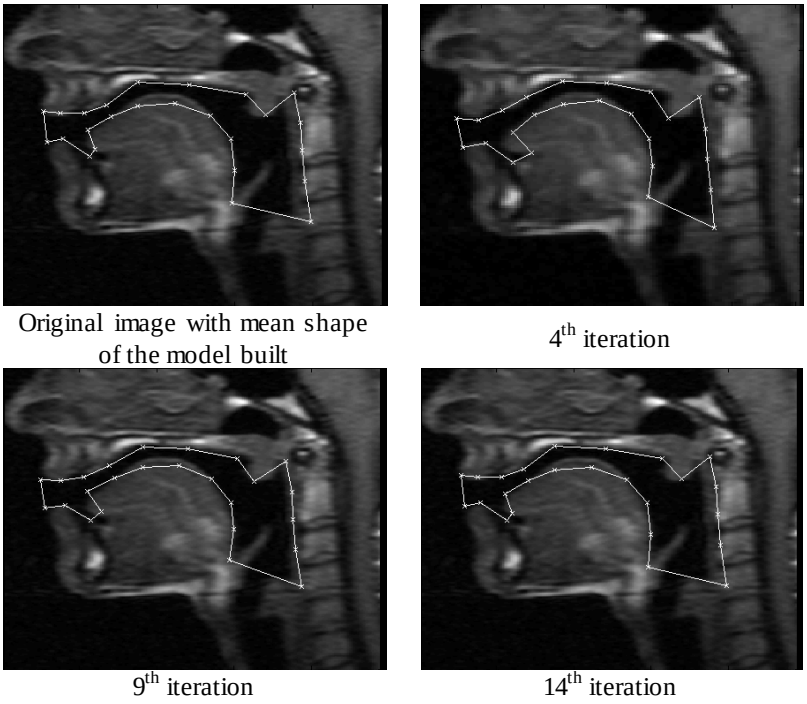


Figure 3.7 – Effects of varying each of the first six modes of variation of the model built for the vocal tract's shape (mean ± 2 standard deviation).

Afterwards, 4 MR images of 4 distinct EP speech sounds, which were not considered in the set of training images used, were segmented by the active shape models built. Figure 3.8 depicts an example the segmentations obtained for one image. Thus, in this figure it is possible to observe some of the iterations of the segmentation process by the active shape model built: it starts with a raw estimation on the localization of the vocal tract in the image (1st iteration), downwards each multiresolution level (4th and 9th iteration) until converges into the final the vocal tract's shape after 14 iterations. This segmentation was obtained considering an active shape model able to explain 95% of all variance of the vocal tract's shape under study and adopting a gray level profile of 7 pixels long, that is considering 3 pixels from each side of the landmark points [Vasconcelos et al. 2010]. Likewise, the segmentation results using this model on all the 4 testing MR images are shown in Figure 3.9.

In Table 3.5, the values of the mean and standard deviation that translate the quality of the segmentation obtained in each testing MR image by the active shape models built are presented. (For a better understand of the data indicated in this table, the models are named as: *Asm_varianceretained_profiledimension*). As it was said earlier, active shape models with gray level profile of dimensions equal to 11 and 19 pixels were also built. However, these active shape models were not able to segment successfully the modeled organ in the testing images. This fail was due to the size of the images considered, namely 256x256 pixels, which is relatively small: during the segmentation process, at each landmark point is considered a segment of 22 (or 38) pixels long in the active search and consequently the model built can easily diverge.



Original image with mean shape of the model built

4^{th} iteration

9^{th} iteration

14^{th} iteration

Figure 3.8 – Testing image with the initial position of the mean shape of the model built overlapped and after 4, 9 and 14 iterations of the segmentation process by an active shape model.

Table 3.5 – Mean and standard deviation (mean ± std) errors of the segmentations obtained from the testing images by the statistical models built.

Models	Image 1	Image 2	Image 3	Image 4
Asm_95_p7	9.99 ± 5.76	9.89 ± 4.43	11.54 ± 6.36	14.23 ± 7.66
Asm_99_p7	9.97 ± 6.27	10.65 ± 3.45	fail	12.25 ± 5.86
Aam_95_5000	4.90 ± 2.42	10.21 ± 5.09	8.98 ± 4.80	9.91 ± 3.95
Aam_99_5000	6.77 ± 3.18	9.73 ± 4.56	8.80 ± 4.88	9.83 ± 4.48
Aam_95_10000	4.94 ± 2.45	10.19 ± 5.07	8.98 ± 4.78	10.56 ± 4.00
Aam_99_10000	4.35 ± 2.30	9.71 ± 4.60	8.80 ± 4.89	10.06 ± 4.58

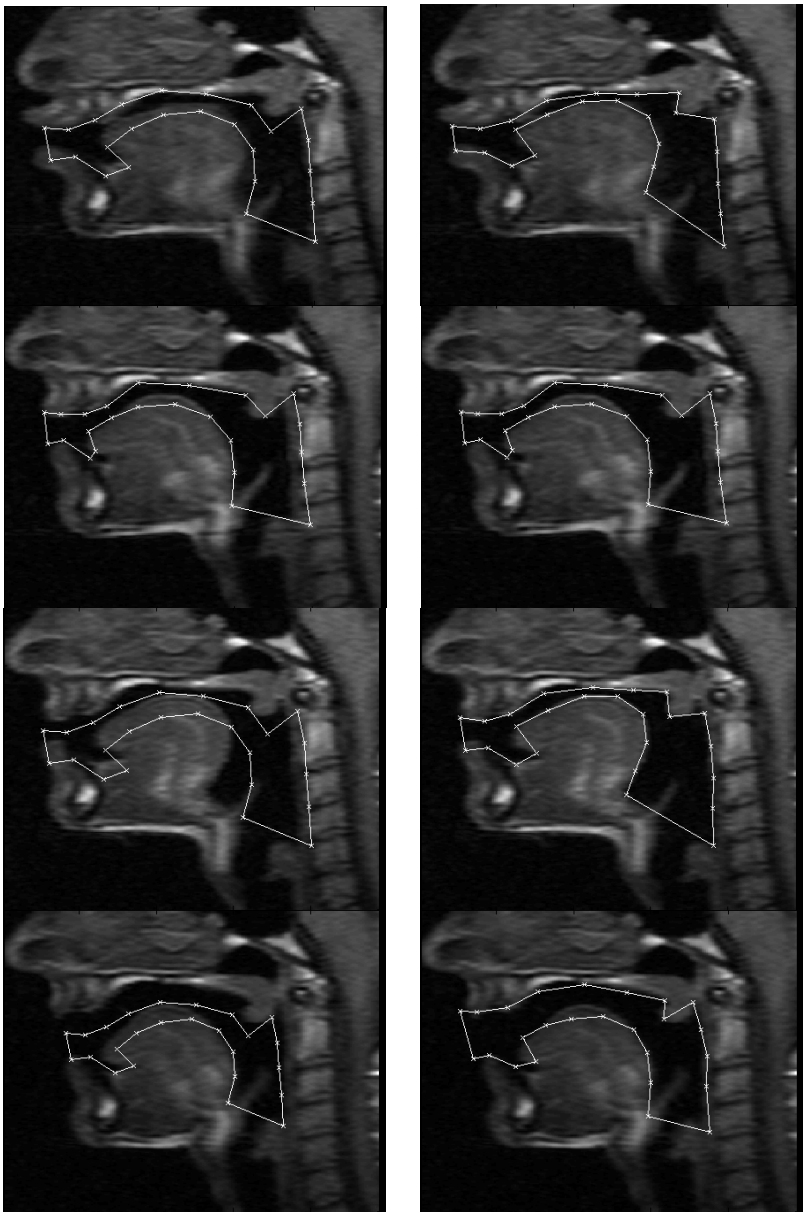


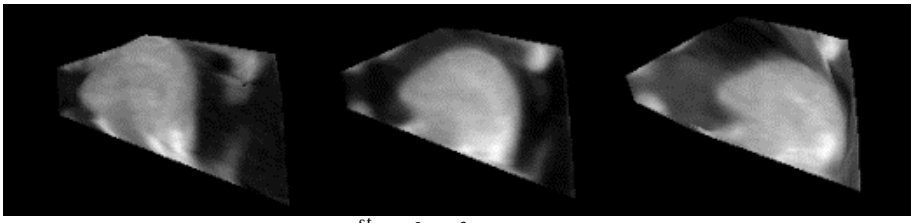
Figure 3.9 – Testing images with the initial position of the mean shape model built overlapped (left) and the final results of the segmentation process by an active shape model (right).

As previously stated, active appearance models can also segment objects modeled in new images. By considering 95% of all the shape's variance and 10000 pixels in the construction of the texture model, 9 modes of shape variation, 17 texture modes and 13 appearance modes were extracted. Additionally, for an active appearance model built using 99% of all the shape variance and considering the same number of pixels, then 15 shape modes, 20 texture modes and 18 appearance modes were obtained.

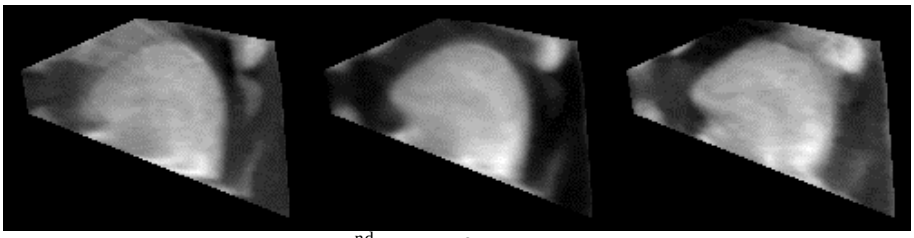
The effects of varying the initial 3 modes of variation of texture and appearance of the active appearance models built are depicted in Figure 3.10 and Figure 3.11. Both figures allow one to become aware that the first mode is associated with tongue's movements from the high front to back positions. Furthermore, one can verify that the second mode of variation is related to the vertical movement of the tongue towards the palate. Finally, the third mode of variation appears to translate the lips' movement together with the tongue's movement to backward. It should be noticed that these modes of variation also contain information about the appearance, meaning that the intensity profiles associated with each structure of the vocal tract are considered.

Figure 3.12 presents an example of the segmentation result using one active appearance models built on a MR testing image. In this figure, it is possible to observe 4 of all the iterations of the active search needed to correctly segment the organ modeled: it starts with a raw estimation on the localization of the vocal tract in the image (1^{st} iteration), downwards each multiresolution level (7^{th} and 12^{th} iteration) until converges into the desired vocal tract' shape after 20 iterations. Similarly, the segmentation results using the same model on the all testing MR images are shown in Figure 3.13. Additionally, the obtained values of the mean and standard deviation that translate the quality of the segmentation obtained in each testing MR image by the active appearance models built are presented in Table 3.5. (Again, for a better understand of the data indicated, the models are named as: *Aam_varianceretained_npixelsused*).

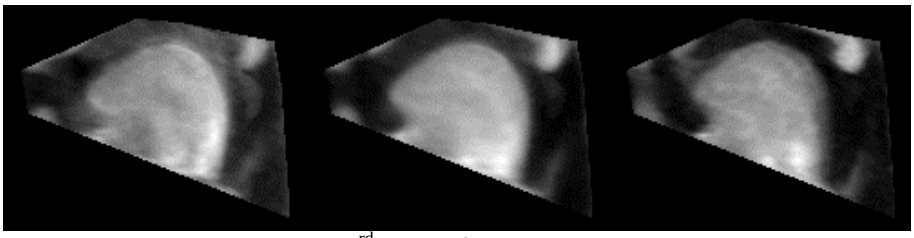
Through an analysis of the data presented in the Table 3.5, one may conclude that the active appearance models obtain superior results than the active shape models. Furthermore, the use of more modes of variation lead to better results when the active appearance models are used, in contrast with the segmentation results obtained by using the active shape models, where the use of more modes of variation (retained percent) not always translated in improved results.



1st mode of variation

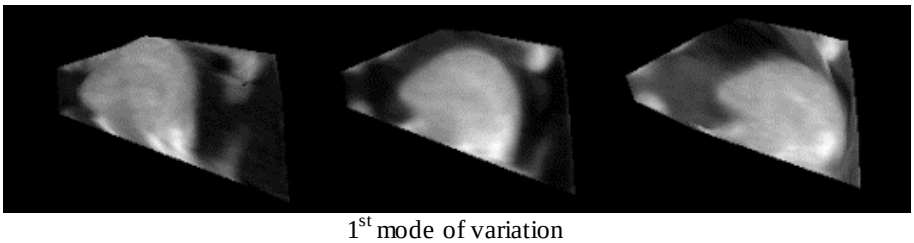


2nd mode of variation

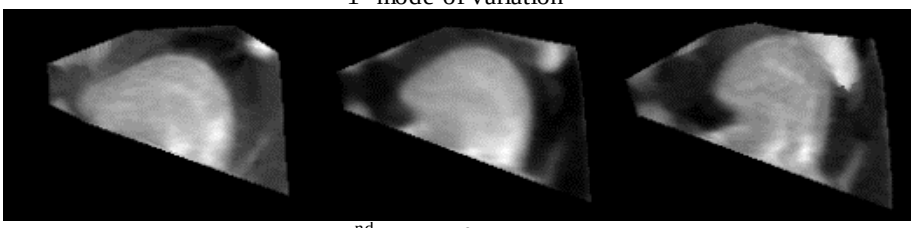


3rd mode of variation

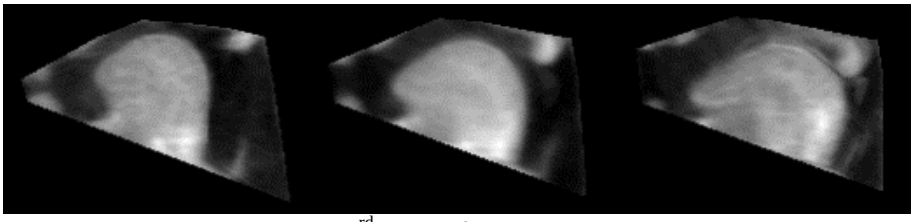
Figure 3.10 – First three modes of texture variation of the active appearance model built for the vocal tract's shape (mean ± 2 standard deviation).



1st mode of variation

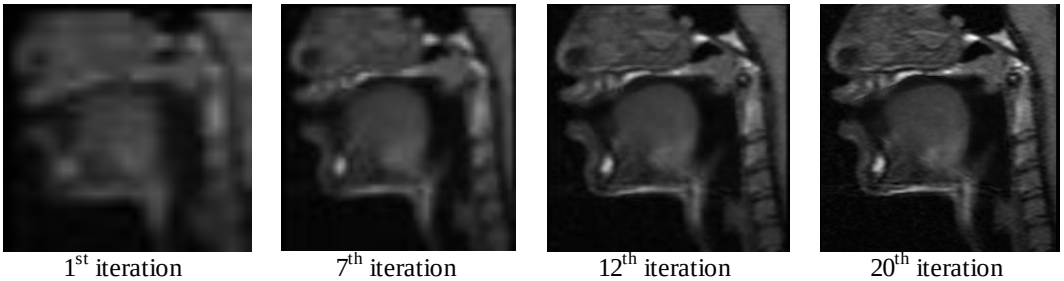


2nd mode of variation



3rd mode of variation

Figure 3.11 – First three modes of appearance variation of the active appearance model built for the vocal tract's shape (mean ± 2 standard deviation).



1st iteration 7th iteration 12th iteration 20th iteration

Figure 3.12 – Results after the 1st, 7th, 12th and 20th iterations of the segmentation process using one active appearance model built for the vocal tract.

Summarizing, statistical deformable models, ASM and AAM, were applied in Magnetic Resonance Images to study the shape of the vocal tract in the articulation of European Portuguese sounds as well as used to segment the vocal tract's shape in new MR images.

While active shape models consider the information around each landmark point of the modeled object, active appearance models use also the gray level information of the object. Consequently, the former type of models tends to be less efficient than the latter, being this information confirmed in this work. Nevertheless, both active shape models and active appearance models obtained remarkable results, either in terms of translating the movements and configurations involved in speech production, as well as in the segmentation of the vocal tract in new images.

The models built could fruitfully extract the main characteristics of the movements of vocal tract from 1.5T MR images. Furthermore, the low mean errors obtained in the segmentation of new MR images, from 4 to 10 pixels for 256x256 pixels images, proved that these models can be accurate and efficient tools to be used towards the automatic study of the vocal tract from magnetic resonance images during speech production.

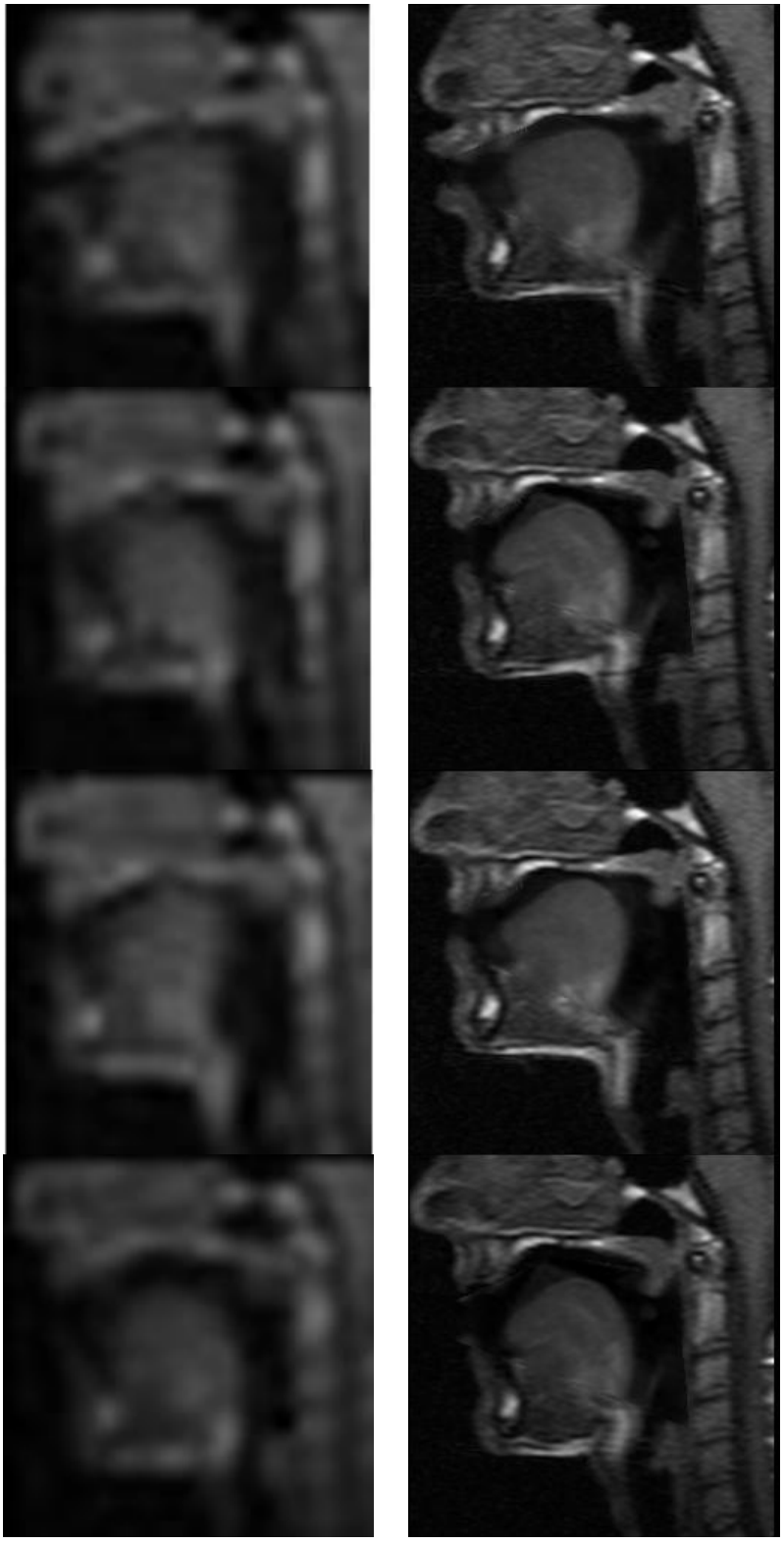
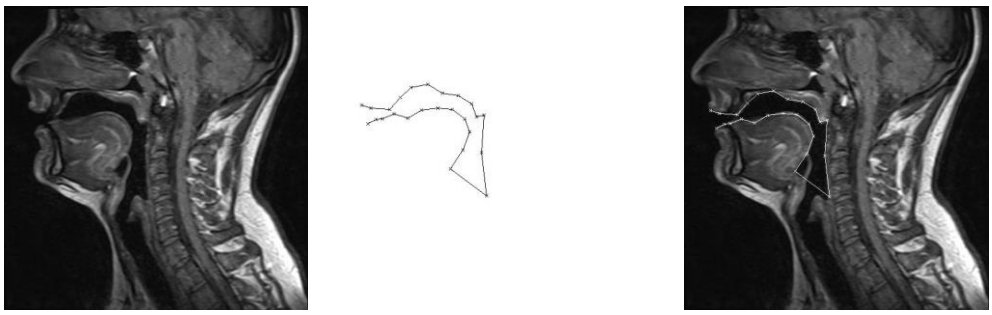


Figure 3.13 – Testing images with the initial position of the mean shape model built overlapped (left) and the final results of the segmentation process obtained by an active appearance model (right).

3.4.5. Vocal Tract Active Models on 3.0T MR Images

The study presented in this section is similar to the previous one, with the difference that images with much higher quality, from the 3.0T sounds dataset, were used in the building of the statistical models of the vocal tract. These models were later used to evaluate the quality of the segmentations in new images [Vasconcelos et al. 2012].

For the construction of the models, 25 sounds have been considered by using 3 MR images per each one from the 3.0T sounds dataset. The localization of the landmarks was consistent with the 1.5T dataset: 25 landmark points were manually extracted anatomical points from the vocal tract articulators by a medical imaging specialist, Figure 3.14.



a) b) c)

Figure 3.14 – a) Training image, b) landmark points selected, c) image labeled with the overlapped landmark points selected.

The sounds considered included the most representative sounds of the EP speech language. Additionally, 2 other MR images for the EP speech sounds /f/, /v/ and /a/ for each subject (making a total of 12 new images) were used to be segmented. The selection of the sounds to be segmented considered that: the associated sounds were easy to sustain, required slight efforts to the subjects and ensured the steadiness of the shape of the vocal tract; also, included the two classes of sounds under study, namely two fricative consonants, one voiced and another one voiceless and one vowel (/a/).

Again, the ASM models built had the same characteristics of the models from the previous section [Vasconcelos et al. 2010]. So, ASMs were built adopting 90%, 95% and 99% of retained variance and profiles of 7, 11 and 15 pixels. Similarly, Active Appearance Models were built adopting equal values of retained variance and the following values of 5000 and 10000 pixels were considered for building the texture model. These parameters were defined based on the previous experience concerning the statistically modeling of vocal tract using these models [Vasconcelos et al. 2008; Vasconcelos et al. 2010; Vasconcelos et al. 2011].

Following the building of the ASMs and AAMs from the training set constituted by 138 images, the models were used to segment the vocal tract in 12 new images. As a stopping criterion of the segmentation process, a maximum of 6 iterations on each resolution level was taken into consideration. Due to the fact that 5 resolution levels were defined and based on the dimensions of the images under study, the criterion means that, from the beginning of the segmentation process to its end, a maximum of 30 iterations can occur. This maximum number of iterations was chosen as a result of the fact that in the experiments done it led to excellent segmentation results.

From Table 3.6, one may observe that the initial 11 modes of variation of the Active Shape Model built are capable of explaining 90% of all variance of the vocal tract. Moreover, one may conclude that the first 17 modes provide an explanation for 95% of all variance and the initial 33 modes of variation illustrate 99% of all variance. Once more, these findings clearly indicate the ability of the built ASM to considerably condense the data that is required to represent all configurations that the vocal tract assumes in the image training set.

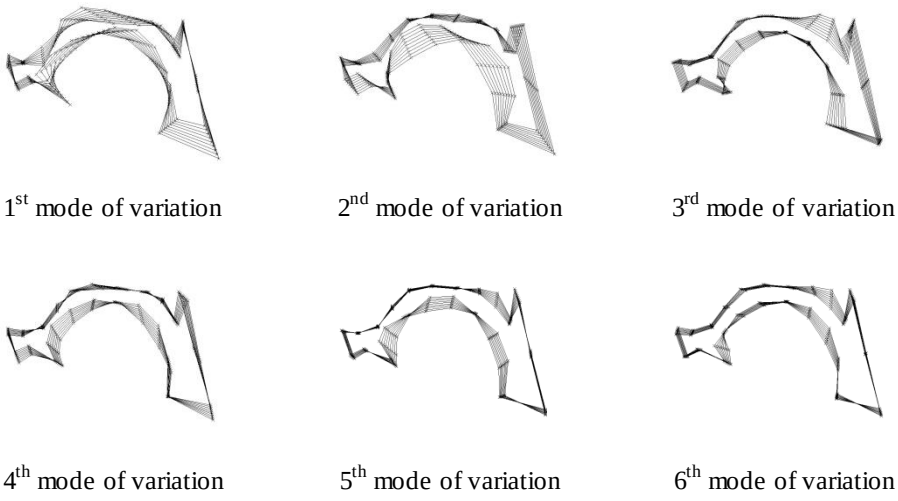
The effects on varying the first 6 modes of variation of the built models are depicted in Figure 3.15. From this figure, one can realize that the first mode is related to the movements of the tongue from the front to the back in the oral cavity associated with the rise of the larynx. With regard to the second mode of variation, it is possible to observe the movements of the tongue from the front-high to the back-down in the oral cavity associated with the lips opening and narrowing. The third mode of variation describes the velum's lowering associated with the enlargement/narrowing of the pharynx cavity and the tongue's tip

movement. The vertical movement of the body of the tongue towards the palate is revealed by the fifth mode of variation. The variations of the sixth mode illustrate the open/close of the lips associated with the vertical movement of the tongue. After this mode of variation, all the remainder modes represent more particular movements, such as the larynx height adjustment, the tongue's tip movement, the opening and closing of the lips, the vertical rise of the tongue's body towards the palate and the pharynx narrowing.

Table 3.6 – Retained percentages along the initial 17 modes of variation of the model built for the vocal tract.

Mode of variation	Retained Percentage	Cumulative Retained Percentage
λ_1	40.893	40.893
λ_2	16.348	57.241
λ_3	8.065	65.306
λ_4	7.404	72.710
λ_5	4.595	77.305
λ_6	3.920	81.225
λ_7	2.515	83.740
λ_8	2.115	85.855
λ_9	1.703	87.558
λ_{10}	1.397	88.955
λ_{11}	1.296	90.251
λ_{12}	1.108	91.359
λ_{13}	1.021	92.380
λ_{14}	0.787	93.167
λ_{15}	0.677	93.844
λ_{16}	0.632	94.476
λ_{17}	0.562	95.038

After the analysis on the ability of the built statistical models to render the real behavior of the vocal tract during the production of EP language sounds, 12 new MR images of the 3 distinct EP speech sounds previously selected (/f/, /v/ and /a/), i.e. of images not included in the used training image set, were automatically segmented by the same models.

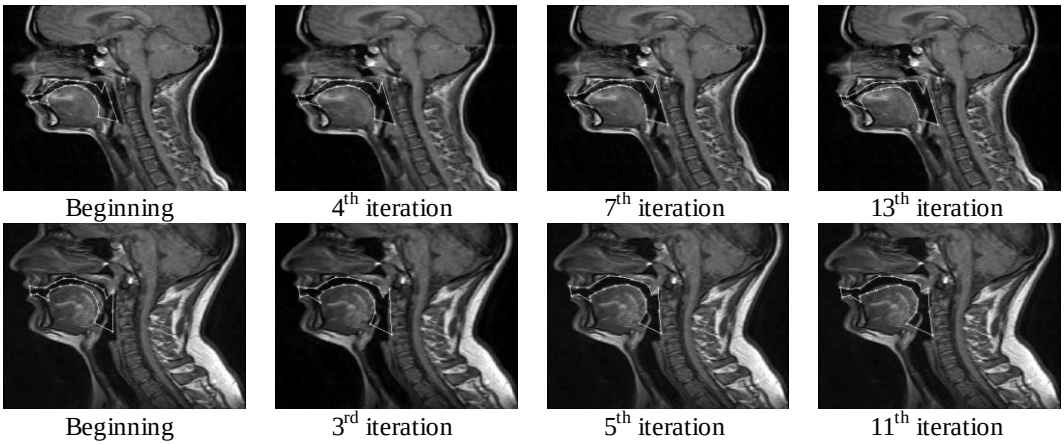


1st mode of variation 2nd mode of variation 3rd mode of variation

4th mode of variation 5th mode of variation 6th mode of variation

Figure 3.15 – Effect on the vocal tract by varying (±2standard deviation) each of the first six modes of variation of the model built.

In Figure 3.16, one MR image of each subject articulating the EP speech sound /f/ is presented as well as the evolution of the correspondent segmentation by the active shape model built: the segmentation begins with a rough estimate for the vocal tract in the input image and then deforms it towards the desired segmentation. Analogously, the segmentation results obtained by using this model on other 4 new MR images are presented in Figure 3.17, where the first two images concerns to one subject and last image to the other subject articulating the EP speech sounds /v/ and /a/, respectively.



Beginning 4th iteration 7th iteration 13th iteration

Beginning 3rd iteration 5th iteration 11th iteration

Figure 3.16 – Test image of female (top row) and male (bottom row) subjects overlapped with the mean shape model built and after some iterations of the segmentation process of the active shape model built.

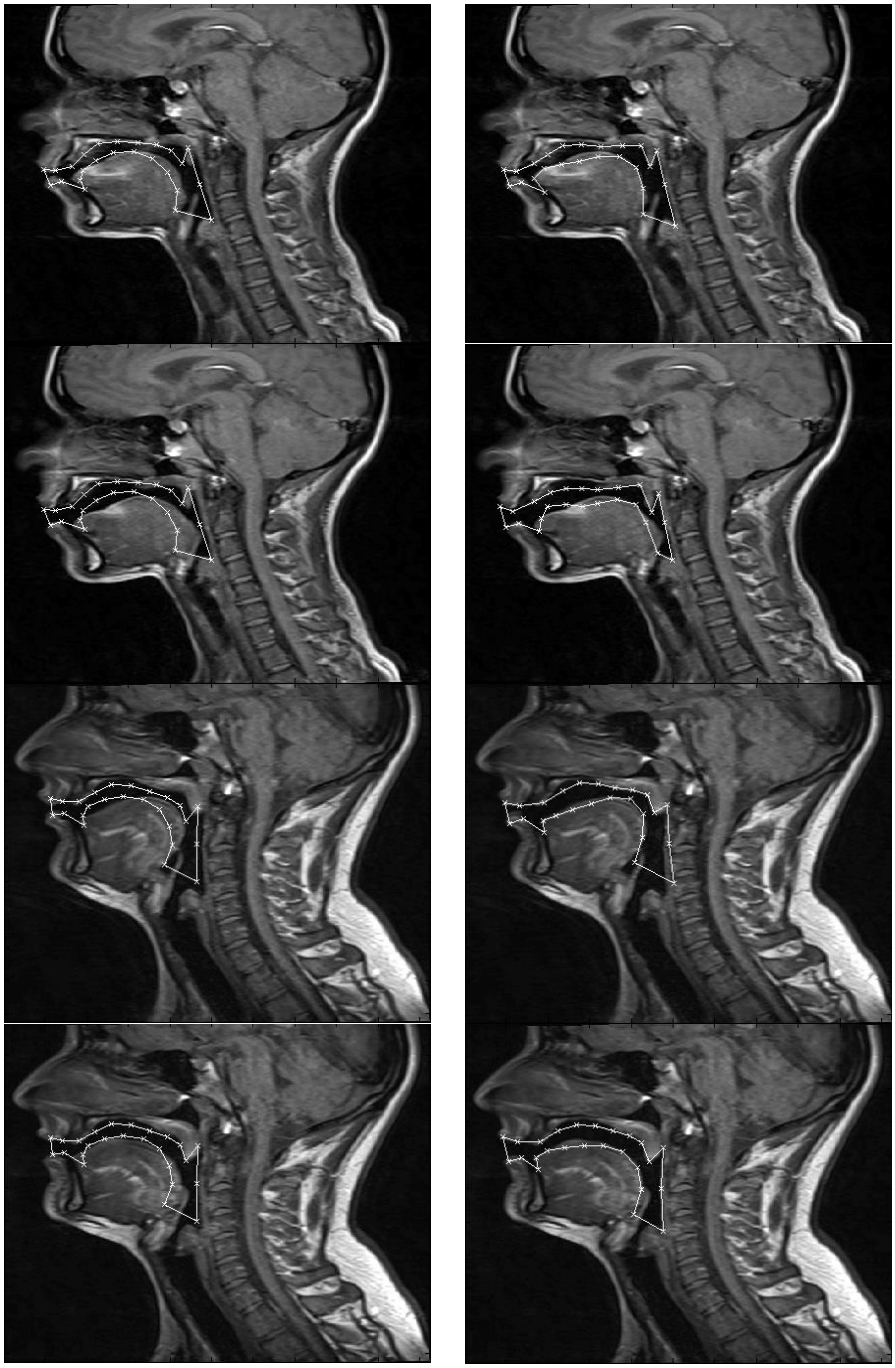


Figure 3.17 – Four test images overlapped with the mean shape model built (left) and after the conclusion of the segmentation process by the active shape model built (right).

The segmentation results depicted in Figure 3.16 and Figure 3.17 were obtained considering an ASM capable of explaining 90% of all variance of the vocal tract under study and adopting a gray level profile length of 11 pixels, that is by considering 5 pixels from each side of the landmark points.

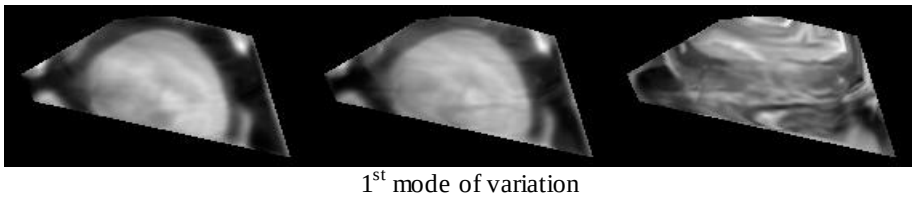
In Table 3.7, the values of the mean and standard deviation that reflect the quality of the segmentations obtained by the active shape models built in each testing MR image are indicated. (As previously adopted, the models are named as *Asm_varianceretained_profiledimension* and cases of segmentation failures are indicated by a dash.) The results concerning the built ASMs considering 99% of all variance were not included in this table, since the models were not able to successfully segment the modeled organ in most of the testing images. This failure is precisely due to the percentage of retained variance used, 99% that led to an extremely rigid model and, because of that, had a very low ability to be adapted to new configurations.

Table 3.7 – Mean and standard deviation (mean ± std) errors of the shapes segmented by the deformable models built.

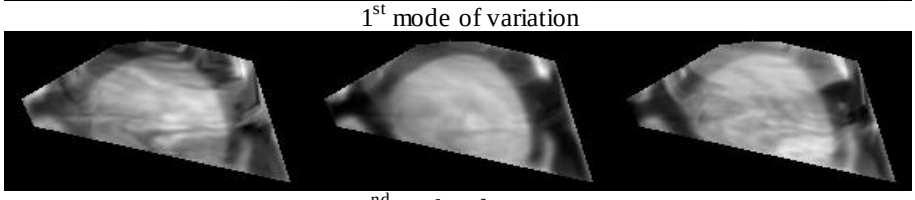
Female subject						
Models	Image 1	Image 2	Image 3	Image 4	Image 5	Image 6
Asm_90_p7	8.99 ± 5.45	8.05 ± 4.92	16.02 ± 14.93	10.78 ± 7.23	13.39 ± 7.65	12.61 ± 7.01
Asm_95_p7	10.40 ± 5.77	9.23 ± 5.92	14.63 ± 8.16	11.07 ± 9.80	14.29 ± 8.70	13.21 ± 8.13
Asm_90_p11	7.50 ± 4.80	7.25 ± 4.42	16.93 ± 14.29	8.70 ± 4.46	17.29 ± 10.62	14.49 ± 8.33
Asm_95_p11	9.89 ± 6.11	10.42 ± 7.48	17.72 ± 14.40	8.70 ± 5.11	-	-
Asm_90_p15	8.28 ± 4.41	8.29 ± 3.44	16.77 ± 15.50	8.38 ± 4.68	16.54 ± 8.05	14.34 ± 8.17
Asm_95_p15	8.29 ± 4.56	8.19 ± 3.78	16.40 ± 15.79	8.67 ± 4.29	16.40 ± 8.73	14.19 ± 8.41
Aam_90_5000	6.75 ± 4.09	7.81 ± 4.74	13.61 ± 15.67	9.37 ± 6.07	9.54 ± 8.36	9.28 ± 8.59
Aam_95_5000	-	6.87 ± 5.89	13.53 ± 15.06	-	8.89 ± 6.53	-
Aam_90_10000	7.04 ± 4.55	7.93 ± 4.76	13.16 ± 15.84	9.05 ± 5.91	9.54 ± 8.50	9.42 ± 8.67
Aam_95_10000	6.43 ± 5.21	6.92 ± 4.91	13.10 ± 14.72	9.63 ± 6.38	8.66 ± 5.15	5.34 ± 2.82
Male subject						
Models	Image 1	Image 2	Image 3	Image 4	Image 5	Image 6
Asm_90_p7	9.35 ± 5.18	9.23 ± 7.31	11.11 ± 7.48	15.35 ± 11.34	7.68 ± 3.53	11.05 ± 6.39
Asm_95_p7	8.93 ± 6.21	10.90 ± 8.48	14.92 ± 8.23	10.20 ± 6.57	-	11.02 ± 9.10
Asm_90_p11	6.51 ± 3.12	6.83 ± 4.12	12.81 ± 7.41	9.80 ± 7.50	7.83 ± 4.65	9.66 ± 4.73
Asm_95_p11	9.08 ± 4.55	8.71 ± 5.71	13.87 ± 7.77	9.65 ± 6.09	8.13 ± 4.53	10.30 ± 5.55
Asm_90_p15	6.53 ± 3.85	9.25 ± 4.54	11.75 ± 6.86	10.33 ± 5.55	8.56 ± 5.91	10.11 ± 7.01
Asm_95_p15	6.25 ± 4.09	8.84 ± 5.07	11.59 ± 7.05	10.46 ± 5.36	8.47 ± 6.11	9.94 ± 7.36
Aam_90_5000	6.75 ± 6.84	11.22 ± 7.13	11.61 ± 7.23	10.05 ± 5.65	7.62 ± 5.68	8.81 ± 5.51
Aam_95_5000	5.06 ± 4.40	10.05 ± 7.29	9.28 ± 6.52	5.32 ± 2.98	-	5.32 ± 4.45
Aam_90_10000	6.93 ± 7.06	11.82 ± 7.47	11.99 ± 7.46	10.37 ± 5.93	7.78 ± 5.78	8.24 ± 5.19
Aam_95_10000	4.91 ± 4.19	11.20 ± 7.59	9.79 ± 6.81	7.96 ± 3.97	6.19 ± 4.81	4.97 ± 3.55

As aforementioned, active appearance models are also proficient in modeling objects in images and to segment the modeled objects into new images.

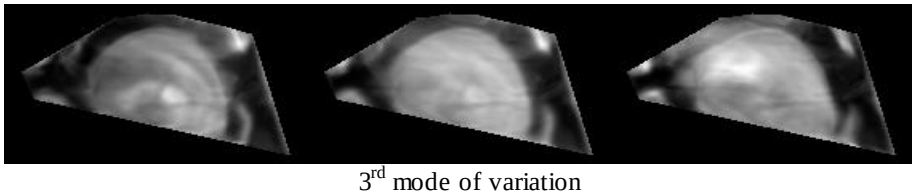
Texture and appearance modes of variation are more difficult to analyze because some motion artifacts ("blur effect") are presented as a result of some inconsistencies of the female subject to sustain the sound, and also because of the inter-subjects differences of vocal tract morphologies. The effects of varying the initial 3 modes of variation in terms of texture and appearance of one of the active appearance models built are depicted in Figure 3.18 and Figure 3.19, respectively. From these figures, it is possible to observe a noticeably number of movements, which are mostly related to the tongue. The first mode of texture depicts the movement of the lower lips and tongue's enlargement in the oral cavity. Whereby, the second mode of variation describes the tongue's tip movement to the alveolar region, and the same movement is observed in association with a backward movement of the tongue in the third mode of variation. On the other side, the first mode of variation of appearance describes the tongue's enlargement in vertical and horizontal directions in the oral cavity. On the other hand, the variation of the second mode demonstrates the forward and backward movements of the tongue associated with the rise of the larynx. Finally, the third mode of variation depicts the forward and backward movements of the tongue in direction to the palate. These results were obtained considering an AAM capable of explaining 95% of all variance of the vocal tract under study and using 10000 pixels in the construction of the texture model.



1^{st} mode of variation

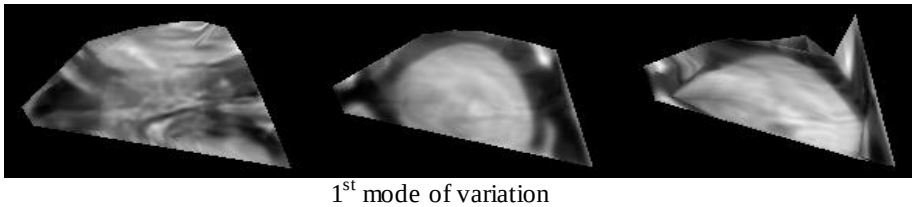


2^{nd} mode of variation

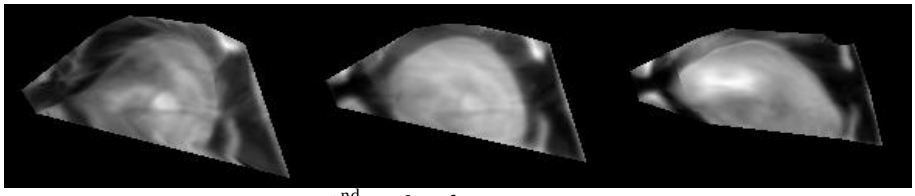


3^{rd} mode of variation

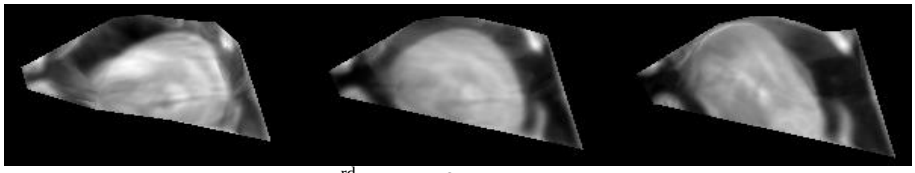
Figure 3.18 – Influence of the first 3 modes of texture variation of the active appearance model built (mean ± 2 standard deviation).



1st mode of variation



2nd mode of variation



3rd mode of variation

Figure 3.19 – Influence of the first 3 modes of appearance variation of the active appearance model built (mean ± 2 standard deviation).

Figure 3.20 presents the segmentation result obtained using one of the active appearance models built on one testing MR image of each subject articulating the consonant /f/. In this figure, one may observe the evolution of the segmentation process through the same active appearance model: the process begins with a rough estimate of the vocal tract in the input image and then deforms it into the final vocal tract configuration. Similarly, the segmentation results obtained by using the model on other 4 testing MR images are depicted in Figure 3.21, where the first two images concerns to the female subject and the last images to the male subject during the articulation of the EP speech sounds /v/ and /a/, respectively. Additionally, the values obtained for the mean and standard deviation in order to translate the quality of the segmentation obtained in each testing MR image by the active appearance models built are included in Table 3.7. (Again, for a clearer understanding of the data indicated, the models have been named as *Aam_varianceretained_npixelsused* and cases of segmentation failures are indicated by a dash.) Similarity as had occurred with the active shape models used, the active appearance models built considering 99% of all variance were not able to successfully segment the modeled organ in most of the testing images and, hence, their results were not included in Table 3.7.

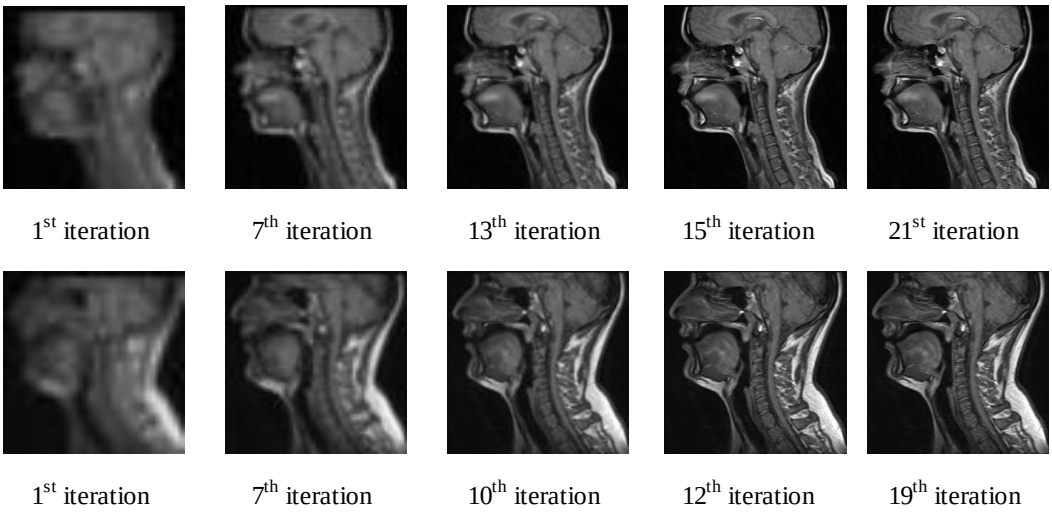


Figure 3.20 – Segmentation process of two test images by the active appearance model built for the vocal tract.

Through the analysis of the data presented in Table 3.7, one may conclude that in comparison to the active shape models, the active appearance models obtained better results, in other words, inferior errors of segmentation. Furthermore, it is possible to realize that the use of more modes of variation not always assure the best results. While ASMs presented enhanced performance when 90% of all variance was addressed, AAMs addressing 95% of all variance had a superior performance when compared with the ones attaining 90% of the variance. Another significant result is that the use of 99% of modes regarding all variance translates in an extraordinary rigid model that it is not capable of be adapted to different configurations, and consequently leading to fail in the segmentation of new images.

The experimental findings are also depicted in Figure 3.22 and Figure 3.23, from which one may verify that the active appearance models performed better than the active shape models. The mean errors obtained for the female subject by the active shape models varied from 7.25 (*Asm_90_p11*-Image 2) to 17.72 (*Asm_95_p11*-Image 3) pixels, 2 situations had occurred in which the segmentation failed. In the other hand, the mean errors obtained by the active appearance models vary from 5.34 (*Aam_95_10000*-Image 6) to 13.63 (*Aam_90_5000*-Image 3) pixels, and 3 unsuccessfully segmentation had occurred. The mean errors obtained for the male subject using the active shape models

varied from 6.25 (*Asm_95_p15*-Image 1) to 15.35 (*Asm_90_p7*-Image 4) pixels, and one unsuccessful case had occurred; while using the active appearance models, the mean errors varied from 4.91 (*Aam_95_10000*-Image 1) to 11.99 (*Aam_90_10000*-Image 3) pixels and the model failed to successfully segment one image.

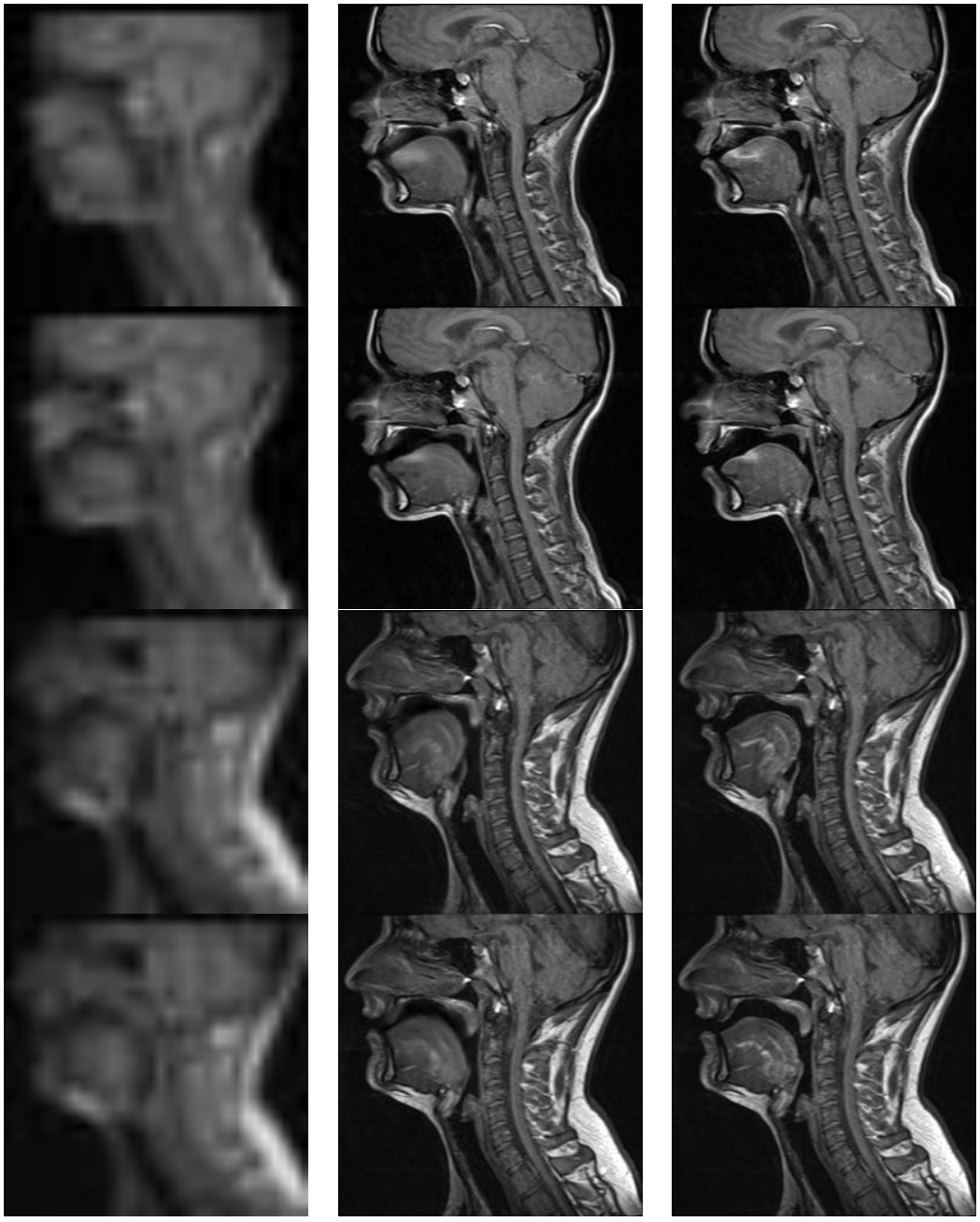


Figure 3.21 – Four test images overlapped with the mean shape model built (left), final results of the segmentation process by the active appearance model built (middle) and correspondent original images (right).

To summarize, 25 out of 30 possible EP speech sounds were modeled for two subjects, being used three measurements (slices) for each sound. Thus, using a training image set of 138 MR images, with more efficient and accurate models than the ones built so far could achieve, as was verified by the experimental findings obtained. Moreover, the images studied were acquired by a 3.0T MR system and, with the higher signal-to-noise ratio and resolution, it was expected that better segmentation results can be obtained when compared to the ones achieved in 1.5T MR images, from the previous section. Indeed, in the previous experiment mean errors rounding 10 pixels were achieved when 256 x 256 pixels 1.5T MR images were used, whilst the segmentation results using the 3.0T MR images led to similar mean errors but in double sized images (512 x 512 pixels).

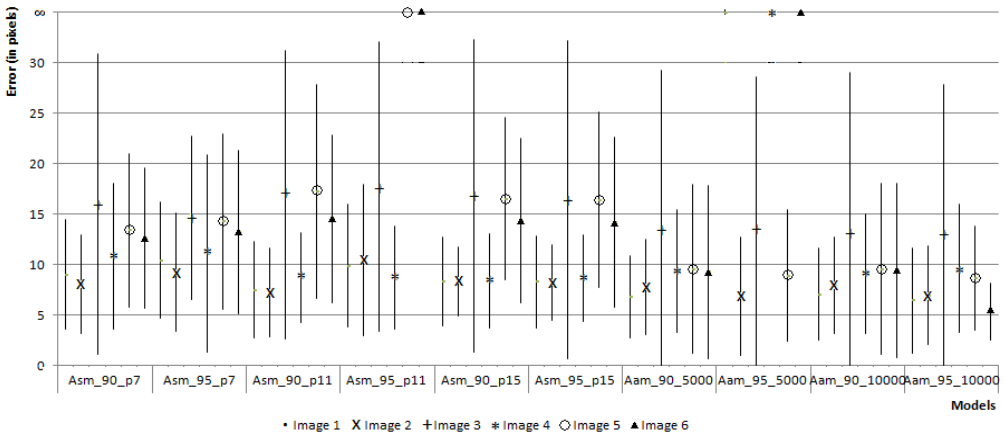


Figure 3.22 – Mean errors (in pixels) and standard deviations of the segmentations obtained by the deformable models built for the vocal tract of the female subject.

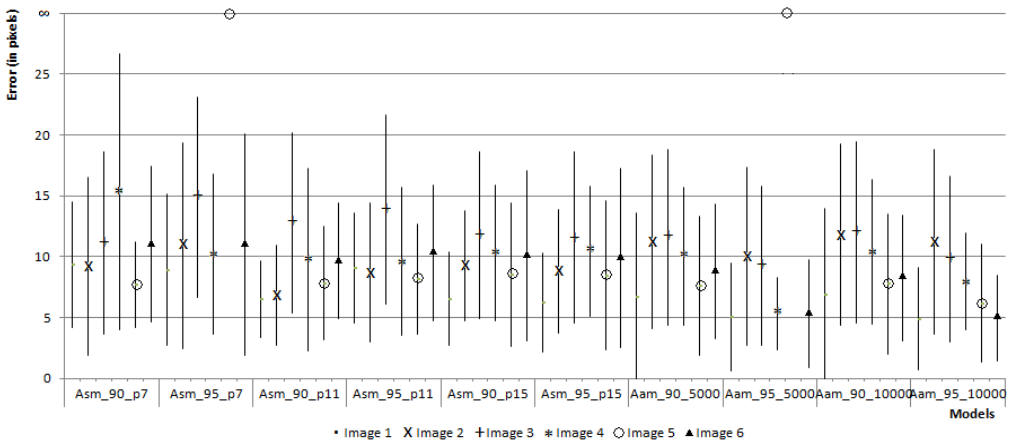


Figure 3.23 – Mean errors (in pixels) and standard deviations of the segmentations obtained by the deformable models built for the vocal tract of the male subject.

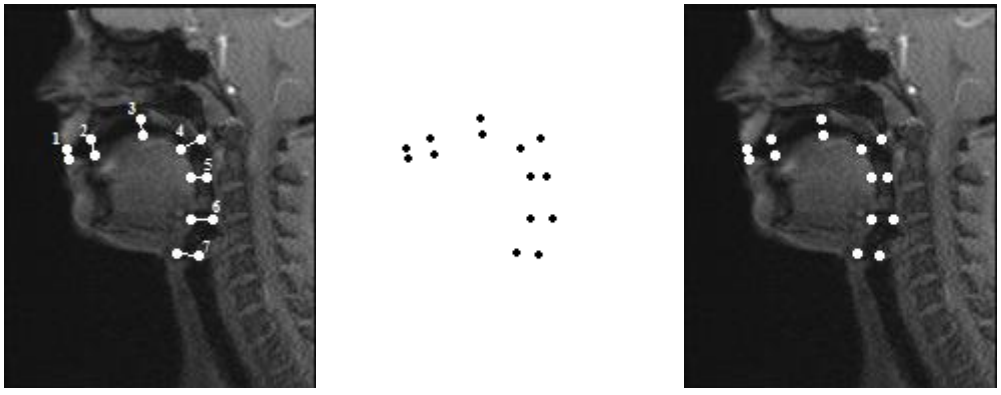
3.4.6. Application Example

The aforementioned models proved to be capable of successfully segment vocal tract articulators in new images. Thus, a use case of the importance of the previous results is here presented [Ventura, Vasconcelos, et al. 2011].

The 3.0T Sequence Dataset is composed by a total of 400 midsagittal MR images, more specifically 100 images for two sequences and for two subjects. For a proper speech articulation assessment, obtained from the quantification of seven articulatory parameters, it was necessary to label the following pairs of landmark points, see Figure 3.24:

- Lips aperture (1);
- Tongue tip constrict location (2);
- Tongue body contrict location (3);
- Velic aperture (4);
- Pharynx width (5);
- Epiglottis distance (6);
- Glottal aperture (7).

Giving a total of 14 landmark points to extract, for each of the 400 images.



a) b) c)

Figure 3.24 – a) Landmark points positions, b) landmark points selected, c) image labeled with the overlapped landmark points selected.

In order to compare the speech articulatory measurements of this large set of images it was necessary to manually annotate each of the 400 images, which is an obviously time consuming task and subsequently highly subjected to errors.

The construction of active vocal tract appearance models, with only 20 images manually annotated per sequence, allowed drastically improving this time consuming step by automatically segment the other images of the sequence, and further obtain the corresponding trajectory distances.

3.5. Discussion and Conclusion

Along this chapter, the automatic study of the vocal tract from 1.5T and 3.0T MR images was assessed through the application of statistical deformable models, namely active shape models and active appearance models. The primary goal consisted on the analysis of the vocal tract during the articulation of European Portuguese sounds, followed by the evaluation of the results concerning the automatic segmentation of the modeled vocal tract in new images.

From the experimental results obtained, one may conclude that the statistical deformable models built are capable of efficiently characterizing the behavior of the vocal tract modeled from the MR images studied. In fact, the modes of variation of the models built could provide an explanation of the actual actions involved in the EP speech sounds considered, such as: the movement of the tongue in the oral cavity, the lip movements or the approximation of the tip of the tongue to the alveolar region. Additionally, it has been verified that the modeling performed could reduce the data set needed to characterize all variations of the shape of the vocal tract during the production of the EP speech sounds.

These models have also revealed that they could easily be used to reconstruct the shape of the vocal tract in the articulation of speech sound. For example, EP speech sounds such as the vowels [i] and [o] or the consonant [j] may be obtained through the combination of just two variation modes, whist the vowel [u] required a combination of the first ten modes of variation to be

successfully reproduced. Also, as a result of the assessment carried out on the reconstructions obtained through the use of the statistical deformable model built, one could analytically prove their elevated quality as all mean errors were inferior to 9 pixels.

Prior to this study, it was believed that EP speech vowels were the easiest sounds to be reproduced as the air flows through the vocal tract without any obstruction. However, the sound that was the most difficult to be successfully reconstructed was the vowel ([u]), thus indicating that it is morphologically more complex to reconstruct this vowel by using the model built.

While active shape models consider the information around each landmark point of the modeled object, active appearance models use also the gray level information of the object. Consequently, the former type of models tends to be less efficient than the latter, being this information confirmed in this work. Nevertheless, both active shape models and active appearance models obtained remarkable results, either in terms of translating the movements and configurations involved in speech production, as well as in the segmentation of the vocal tract in new images.

From the experimental results obtained, one may state that the point distribution model built can fruitfully extract the main characteristics of the movements of vocal tract from magnetic resonance images. Furthermore, one can verify that the active shape models and the active appearance models can be used to segment the modeled vocal tract in new MR images in a successful and automatically manner. Therefore, the models built can be accurate and efficient tools to be used towards the automatic study of the vocal tract from magnetic resonance images during speech production.

One of the premises for acquiring an efficient deformable model, and consequently obtaining good results concerning the segmentation of the modeled object, is extremely related to the quality of the images to be studied. In this chapter, datasets with different image qualities were analyzed, acquired by a 1.5 Tesla and a 3.0 Tesla MR systems. Indeed, for the 1.5T dataset, where 256 x 256 pixels 1.5 Tesla MR images were used, mean errors rounding 10 pixels were

achieved; whilst the segmentation results using the 3.0 Tesla MR images led to similar mean errors but in double sized images, 512 x 512 pixels.

Another major contribution accomplished concerns to the amount of data studied. For the 3.0T sounds dataset, 25 out of 30 possible EP speech sounds were modeled for two subjects, being used three measurements (slices) for each sound. Thus, using a training image set of 138 MR images, with more efficient and accurate models than the ones built so far could achieve, as was verified by the experimental findings obtained.

As a final remark, and after realizing the suitability of this statistical modeling technique to segment the vocal tract in new images, the one presented an use case to prove that these models can, indeed, help imaging experts and speech therapists in this task. By constructing vocal tract active models with the manual annotation of only a one fifth of the dataset, it was possible to rapidly segment the others four fifths of the dataset, instead of manually annotating all images.

To conclude, from the work here described, one should emphasize that the recent MR imaging systems, in particular the 3.0 Tesla, and the use of the adopted statistical modeling technique have made possible the automatically and realist simulation of the vocal tract during speech production as well as the efficient segmentation of vocal tract in new images. Therefore, the assessment of the articulators' positions and movements can be facilitated, contributing, for example to: speech rehabilitation, as a supplementary tool for the therapeutic planning and follow-up for physicians and speech therapists; simulation purposes, namely to recognize and simulate the compensatory movements of the articulators during speech production; and construct improved computational speech models and devices.

4

Silhouette Models

Human motion is one of the most interesting subjects of image analysis due to its promising and important applications in many key fields. The study of human motion can be divided into three different but interconnected steps: the first deals with the segmentation, or identification, of the subject in the images; the second is related to tracking; and finally, the third, in which human motion understanding is performed. Each one of these steps is highly complex and numerous studies have been done to develop methodologies capable of performing such actions, as demonstrated in the second chapter of this Thesis.

This chapter will focus on the first step of human motion analysis, in which the segmentation of a subject is performed. Different methods of image segmentation are reviewed here and applied to different image sequences, from basic methods that model the background to extract the subject in the scene to more complex ones which learn to adapt to changes throughout the image sequences in order to obtain better segmentation results. Since our focus of motion segmentation is a human subject, we present a model that describes the possible silhouettes to be obtained from image sequences with a subject normally walking together with specific landmarks that represent important anatomical points. The model presented can also be used to segment the subject in new images in order to, for instance, further analyze the subject's motion.

The first section of this chapter presents the different image sequences used throughout this study. In the second section, a description of the four different background subtraction models is given and, in the third section, the active silhouette model is explained. The error segmentation measures adopted are described in the fourth section and the details on implementations developed during this work are provided in the fifth section. The sixth section contains the segmentation results obtained with the different methods. Finally, in the seventh section, the results are discussed and conclusions are drawn.

4.1. Image Sequences

Different image sequences were used to evaluate the aforementioned models. The majority of the image sequences belong to widely known datasets that were referred to previously in the second chapter of this Thesis.

A brief description of the image quality, number and position of the subjects as well as the surrounding environment where the sequences were acquired are given in the following subsections.

4.1.1. NADA

The first image sequence used belongs to the NADA database indicated in [Schuldt et al. 2004]. In this sequence a male subject is walking parallel to the camera view in an outdoor environment with a homogeneous background. It is a 22-second video, with 25 frames per second and with image resolution of 160x120 pixels. In Figure 4.1 some examples of images from this sequence are illustrated.

Figure 4.1 – Examples of images extracted from the NADA image sequence.

4.1.2. CASIA-A

The second image sequence used belongs to the CASIA-A dataset [L. Wang, Tan, et al. 2003], the former NLPR gait database. One male subject walks parallel to the camera in an outdoor environment more complex than the previous sequence. The sequence is provided in a 27-second video, with 25 frames per second and with image resolution of 320x240 pixels. Examples of images from this sequence are shown in Figure 4.2.

Figure 4.2 – Examples of images extracted from the CASIA-A image sequence.

4.1.3. CAVIAR

The other image sequence used belongs to the CAVIAR Test Case Scenarios [Fisher et al. 2003], which was taken from a Shopping Center in Portugal and shows a frontal view of the scenario with people walking along a corridor. In this sequence a female subject is walking parallel to the camera view in an outdoor environment with a heterogeneous background. A 16-second video was considered in this study, with 15 frames per second and with image resolution of 320x240 pixels. In Figure 4.3 examples of images from this sequence are shown.

Figure 4.3 – Examples of images extracted from the CAVIAR image sequence.

4.1.4. CASIA-B

The final image sequences considered in this work belong to the CASIA Gait Database (CASIA-B) [Yu et al. 2006]. Information was used from 11 subjects walking in four different directions (0º, 36º, 54º and 90º) in relation to the image camera in an indoor environment, such as the portrait in Figure 4.4. All image sequences are stored as video files encoded with MJPEG, a frame rate of 25 fps and a frame size of 320x240 pixels.

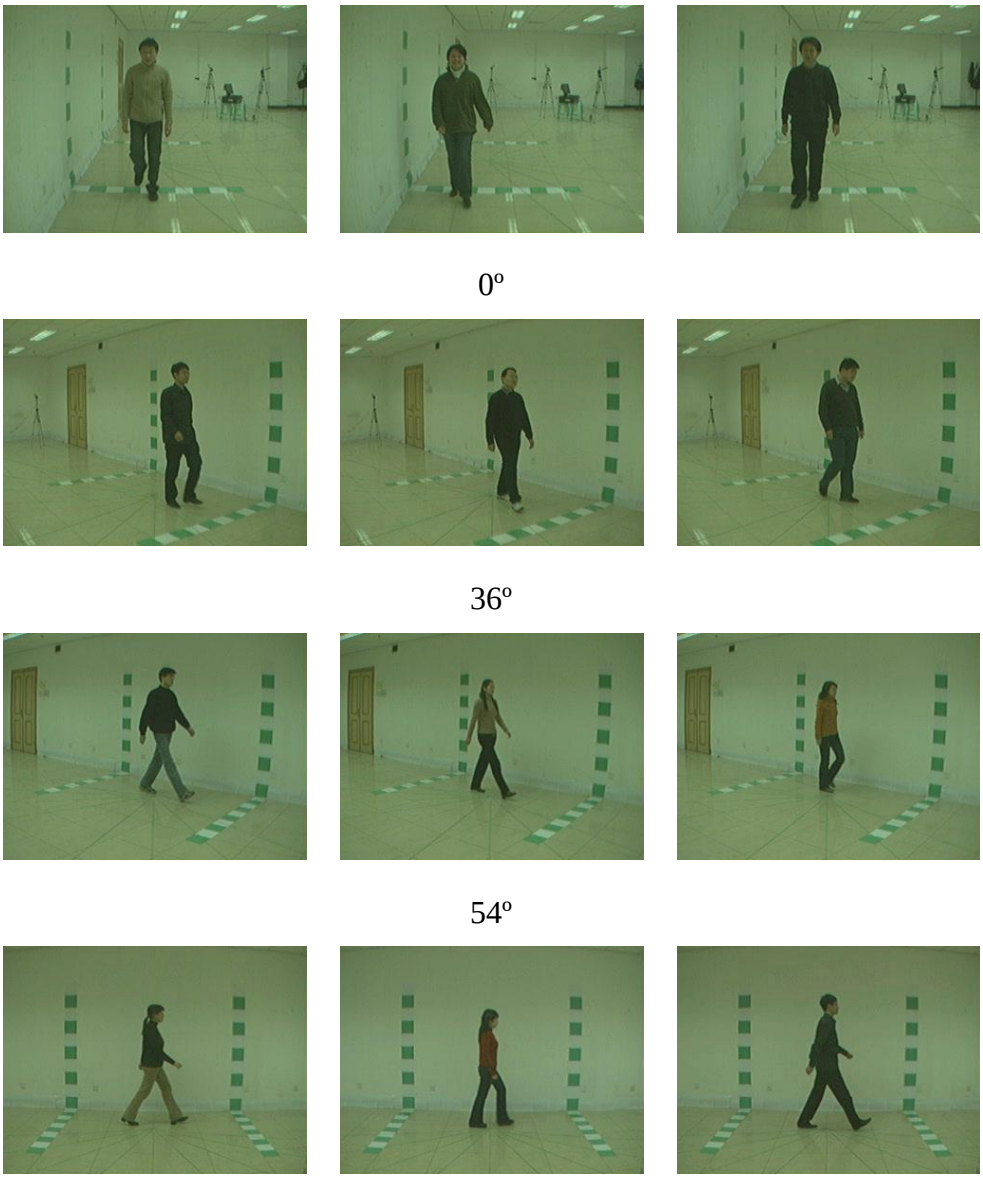


Figure 4.4 – Examples of images extracted from the CASIA-B image sequence, with different subjects and images taken from different views.

4.2. Background Subtraction Models

The use of an edge detection algorithm, by itself, is obviously not sufficient for identifying an object in an image sequence. Thus, more complex models, capable of learning the background and segment moving objects were developed. Four different models were implemented, in the scope of this Thesis, which are described in this section.

4.2.1. Simple Difference Model

Background subtraction is the fundamental method of image segmentation [Piccardi 2004]. It involves the calculi of a reference image followed by the subtraction of each frame of the image sequence from the reference and further threshold of the result.

Here, instead of using the difference between two frames, three frames from different phases of the sequences were used to obtain three difference frames from each other; afterwards, the background image was designed based on all the pixels classified as background from the three threshold difference frames. The foreground is detected by subtracting the present image from the reference background and final threshold of the result.

4.2.2. Running Average Model

Another basic method to obtain the background uses the running average [Wren et al. 1996]. Here, a pixel is classified as background when the pixel value belongs to the corresponding distribution of the background; otherwise, the mean of the distribution is updated. The updated background is then used in the next image. The following equation presents the running average background update method used in this Thesis:

$$\mu_{i+1} = \alpha \times p_i + (1-\alpha)\mu_i, \qquad 4.1$$

where p_i is the pixel value at a given frame i , μ_i is the current average value and the parameter α is the learning rate that defines the influence of the current pixel over the currently estimated background.

4.2.3. Mixture of Gaussians Model

This approach was based on the study of [Stauffer et al. 1999], where the authors introduce a method to model each background pixel by a combination of K Gaussian distributions, with K being a small number between three and five and the different Gaussians are assumed to represent different colors. The weight parameters of the combination represent the time proportions that these colors remain in the scene, so background components will be the ones with the highest probable colors. In other words, the Gaussians of the combination that correspond to background colors are determined based on the persistence and the variance of each of the Gaussians. To allow the model to adapt to changes in illumination, an update scheme is applied based on selective updating. Finally, pixel values that do not fit the background distributions are considered foreground until there is a Gaussian that includes them with sufficient evidence supporting it. According to the authors, only two parameters are necessary to set the system: one defining the time constant that determines the speed at which the distribution's parameters change and another that indicates the minimum portion of data that should account for the background.

To overcome some limitations of the previous approach, such as slow learning in the initial frames, especially in complex environments and difficulty in distinguishing moving shadows from moving objects, the study of [KaewTraKulPong et al. 2002] presents a solution to these problems. The authors reinvestigate the update equations and use different equations at different phases in order to allow the system to learn faster. In addition, the authors incorporate a shadow detection scheme to the model.

4.2.4. Foreground Object Detection Model

Another possibility is the use of the method based on Bayes decision theory to detect foreground objects from complex image sequences presented in [Li et al. 2003]. In the first step of this method, non-change pixels in the image stream are filtered out by simple background and temporal differences. Then, the detected changes are separated as pixels belonging to stationary and moving object according to inter-frame changes. After this, the pixels associated with stationary

or moving objects are classified as background or foreground based on the learned statistics of colors through the use of the Bayes decision rule. The foreground objects are segmented by fusing the results from both stationary and motion pixels. At last, the background model is updated.

The reason of choosing this methodology to test in this project is because it showed to work well in complex backgrounds including sequences with variable light conditions and shadows of moving objects.

4.2.5. Human Silhouette Extraction

After applying the four background subtraction models, post-processing was performed in the segmented images in order to obtain only the human silhouette. First, noise was removed from the resulting images by performing erosion, then all object contours from the image were obtained and the object with larger area was considered the human silhouette.

Besides the normal contour points extracted directly from the silhouette, another silhouette contour was obtained with 100 landmarks defined according to the following distribution:

- 45 points from the left side of the contour (equally spaced);
- 45 points from the right side of the contour (equally spaced);
- 10 points between the feet of the subject (equally spaced);

where the separation of the contour is dictated by three fixed landmark points corresponding to the head and feet. The head is considered the highest left contour point from the silhouette and the feet are obtained as the left and right more distant points from the head.

Figure 4.5 shows an example of the silhouette obtained from an image (a), the silhouette contour (b) and the silhouette contour with 100 landmark points (c), where the red * indicates the head and feet.

a) b) c)

Figure 4.5 – A silhouette example a); silhouette contour extracted from a), b); and 100 contour points extracted from the silhouette a).

4.3. Active Silhouette Model

A Point Distribution Model (PDM) expresses the mean shape of the modeled object in addition to the admissible variations in relation to the same mean shape [Cootes, Taylor, et al. 1992]. Also if, in addition to the geometrical information, gray level information is used, then it is possible to build Active Shape Models (ASM) to segment the modeled object in new images. These methods were explained in detail in the previous chapter and taking into account the successful results, one decided to also explore this methodology also for human silhouette modeling.

In this work, the contour together with the anatomical landmark points of the human shape was modeled by an ASM from a set of 2734 images. The images include various shape configurations of different persons walking, namely, information from 11 subjects walking in four different directions (0º, 36º, 54º and 90º) in relation to the image camera, as illustrated in Figure 4.4.

Table 4.1 shows the number of images employed in the construction of the models built, as well as the number of images used to further test their segmentation accuracy.

Table 4.1 – Summarized table of the data used to build and test the ASM.

Direction	Training images	Test images
00	746	73
36	696	52
54	695	65
90	597	25
Total	2734	215

In order to obtain a robust PDM, the images used in the training process ought to adequately represent the variability of the human shape during walking. Moreover, each shape of the silhouette presented in the training set should be described by a group of labeled landmark points conveying important aspects of the body contour. Hence, 100 contour points were chosen to be extracted from the silhouettes available on the image dataset and further manually annotate 13 extra points indicating the anatomical points from the human stick figure, leading to a total of 113 landmark points to represent the human body structure.

The 100 contour points were obtained as explained in the previous section and the 13 extra points correspond to:

- 1 point in the center of the head (1);
- 1 point on each shoulder (2);
- 1 point on each elbow (2);
- 1 point on each hand (2);
- 1 point on the left and right of the hip (2);
- 1 point on each knee (2);
- 1 point on the backside of each foot (2).

The 13 anatomical landmark points corresponding to the stick figure of the human shape were manually extracted from each frame, and the associated coordinates were concatenated with the contour related landmarks.

In all images to be presented in this chapter, the landmark points corresponding to the human contour appear connected by fictitious line segments to enhance their visualization, while the anatomical landmark points appear represented by the “x” sign. Figure 4.6 show examples of some images used with the corresponding extracted landmark points.

Figure 4.6 – Example of landmark points considered in the four directions (0º, 36º, 54º and 90º) to build the model.

The Active Shape Model was built adopting 95% of all object shape variance in the geometrical modeling (i.e. in PDM) and a profile width of 7 pixels for the gray level modeling. As a stopping criterion for the segmentation process, a maximum of 6 iterations for each resolution level was taken into consideration. Hence, due to the fact that 3 resolution levels were defined based on the dimensions of the images under study, a maximum of 18 iterations could be performed. Other Active Shape Models were built to model the human silhouette, with different retained percentages and profile lengths, with the previously referred to being the more robust and presenting the best results and thus explored in detail here.

From Table 4.2 one can observe that the first 11 modes of the shape model built could explain 90% of all shape variance of the silhouette. The first 20 modes explain 95% of all shape variance, and with only 62 modes of variation, it is possible to explain 99% of all shape variance of the silhouette.

Table 4.2 – Retained and cumulative percentage of the modes of variation of the silhouette model.

Mode of variation	Retained %	Cumulative Retained %
λ_1	40.817%	40.817%
λ_2	13.198%	57.014%
λ_3	11.018%	68.032%
λ_4	6.3333%	74.365%
λ_5	4.4633%	78.829%
λ_6	3.1392%	81.968%
...		
λ_{11}	1.0350%	90.289%
...		
λ_{20}	0.2997%	95.231%
...		
λ_{62}	0.0264%	99.002%

Through the observation of the first four modes of variation with more significance of the PDM built, i.e. the modes shown in Figure 4.7, one can foresee the adequacy of the modeling process for characterizing the human shape. For instance, it is well known that the first mode of variation gathers the information on the walking stance of the subjects, while the second and third modes of variation gather information on the direction in which the subject is walking. In contrast, smaller and more specific variations can be seen in the fourth mode.

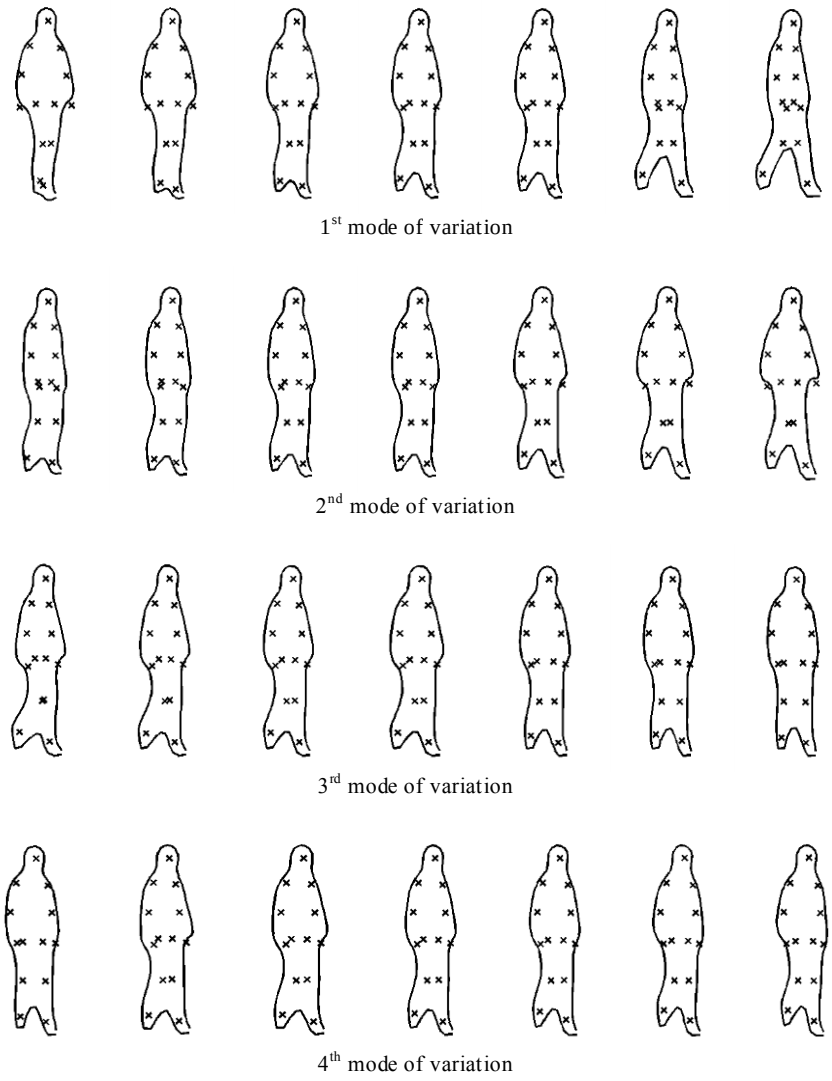


Figure 4.7 – First four modes of variation of the PDM built (mean shape ± 1 std).

4.4. Segmentation Quality Assessment

In order to assess the segmentation quality of the presented methods, in addition to the subjective evaluation from the observation of the segmented images, three objective measures were taken into consideration throughout this work. The first measure, the F-measure, considers the entire image, while the second and third measures, Euclidean and Hausdorff distance, only consider the silhouette contours, or landmark points, instead.

4.4.1. F-measure

While the true positives (TP) provide the number of correctly identified foreground pixels, the true negatives (TN) provide the number of correctly detected background pixels. On the contrary, the false negatives (FN) are pixels wrongly classified as background, whereas false positives (FP) are wrongly classified as foreground. Typical measures for two class problems are: the recall R, which is the ratio between the TP with the number of relevant pixels in the ground truth data (TP+FP); and precision P, which is the ratio between the TP to the total number of pixels (TP+FN). The F-measure combines these two complementary measures with equal weights by calculating:

$$F = \frac{2 \times P \times R}{P + R}. \qquad 4.2$$

4.4.2. Euclidean Distance

The simplest distance to compare shapes is the Euclidean distance. Considering a and b as points of the segmented and ground truth silhouettes, respectively, the Euclidean distance $d(a, b)$ is given by:

$$d(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}, \qquad 4.3$$

where (x_a, y_a) and (x_b, y_b) correspond to the co-ordinates of the landmark points. The Euclidean distance is calculated to each point of the silhouette and the mean and standard deviation of all distances are computed to define the segmentation quality.

Since this measure can only be obtained when the number of points of the segmented silhouette and the ground truth are equal, the next measure was taken into account in order to cope with this drawback.

4.4.3. Hausdorff Distance

The Hausdorff distance has been widely used to compare images [Huttenlocher et al. 1993]. It is the maximum distance of a set to the nearest point in the other set [Rote 1991]. Thus, the Hausdorff distance from a set A to a set B is defined as:

$$h(A,B) = max_{a\in A}\{min_{b\in B}\{d(a,b)\}\}, \qquad 4.4$$

where a and b are points of sets A and B respectively, and $d(A,B)$ is any metric between these points. In this Thesis, one considers $d(A,B)$ as the Euclidean distance between a and b.

4.5. Implementations

The algorithms to create the background subtraction models were implemented using the programming language C++ with the open-source toolkit OpenCV (http://opencv.org), while the active silhouette model was developed using MATLAB (www.mathworks.com), based on the *Active Shape Models software* presented in [Hamarneh 1999].

For the extraction of the 100 landmark points from the silhouette contours, an algorithm in MATLAB was developed to automatically identify and extract the coordinates of the 100 contour points according to the previously described rule in section 4.2.5.

In addition, an implementation for segmentation quality assessment was also developed in MATLAB. Therefore, the three objective measures referred to in the previous section were implemented, namely F-measure, Hausdorff distance and Euclidean distance. For the F-measure, instead of considering the complete image, the bounding box that contained the ground truth silhouette was considered in order to exclude the image area that only contains the background.

4.6. Segmentation Results

In this section, one presents the segmentation results obtained using the background subtraction models as well as the silhouette models.

4.6.1. Background Subtraction Models

For the image sequence NADA, the last 25 images of the sequence in which the subject silhouette was completely present were used to evaluate the four background subtraction segmentation models previously described.

Figure 4.8 shows three of the images belonging to the 25 test images, together with the ground truth silhouette provided with the dataset and the segmentation results obtained using each one of the four background models: simple difference, running average, mixture of Gaussians and foreground object detection models. At first glance of the figure, it is clear that simple difference model has the worst results, as expected, and the foreground object detection has some issues with the last test images, probably due to illumination differences noticed in the right side of the scene. Regarding the running average and the mixture of Gaussians models the subjective evaluation, meaning taking into consideration only the observation of segmentation images obtained, provides quite similar results.

Passing to a more thorough evaluation, namely by an objective evaluation, the mean F-measures, Hausdorff and Euclidean distances for each tested image are present in Figure 4.9, Figure 4.10 and Figure 4.11, respectively. In addition, Table 4.3 presents the mean results of the segmentation errors of all tested images together with its standard deviation.

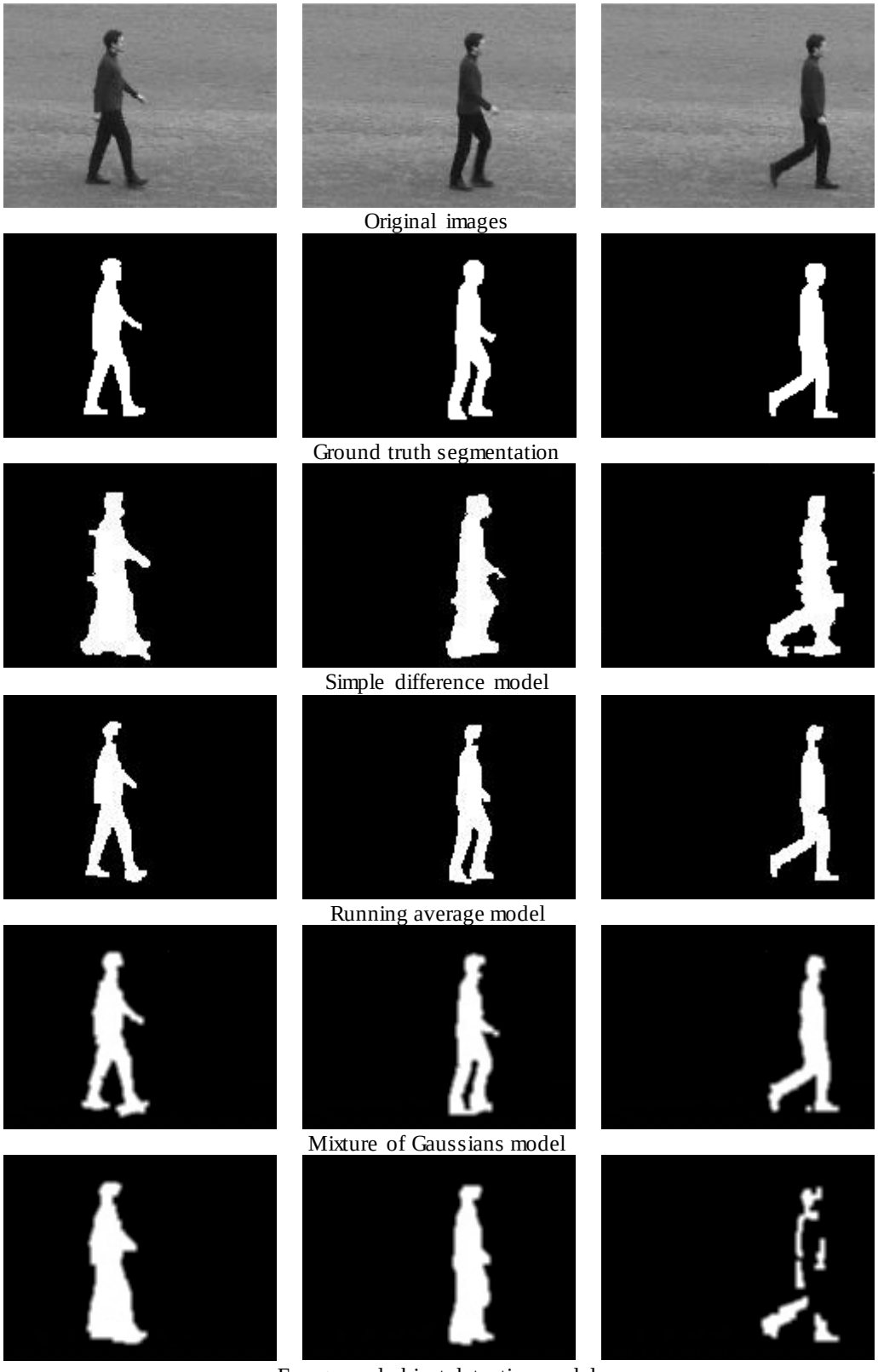


Figure 4.8 – Three images from the NADA image sequence, the respective silhouette ground truth and segmentation results using the different background subtraction models.

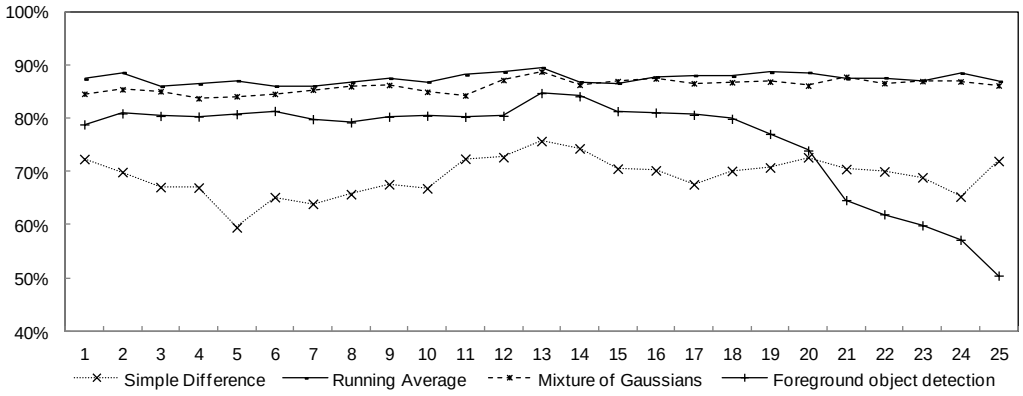


Figure 4.9 – Mean F-measures obtained using the different studied segmentation models for each test image of the NADA image sequence.

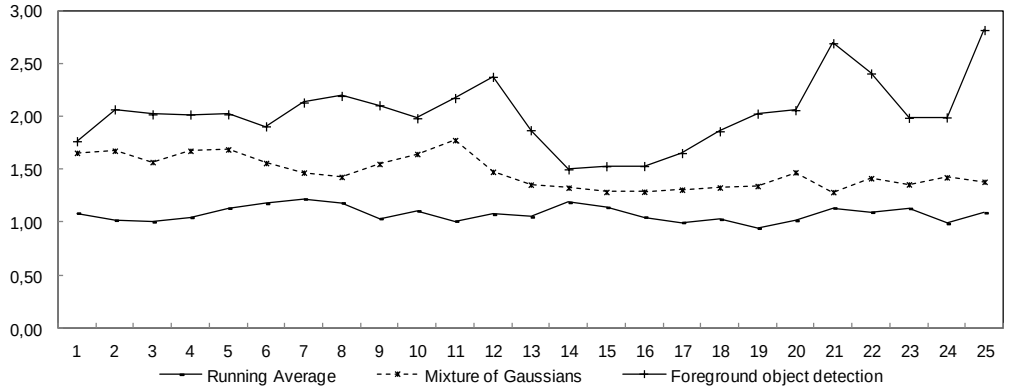


Figure 4.10 – Mean Hausdorff distances obtained using the different studied segmentation models for each test image of the NADA image sequence.

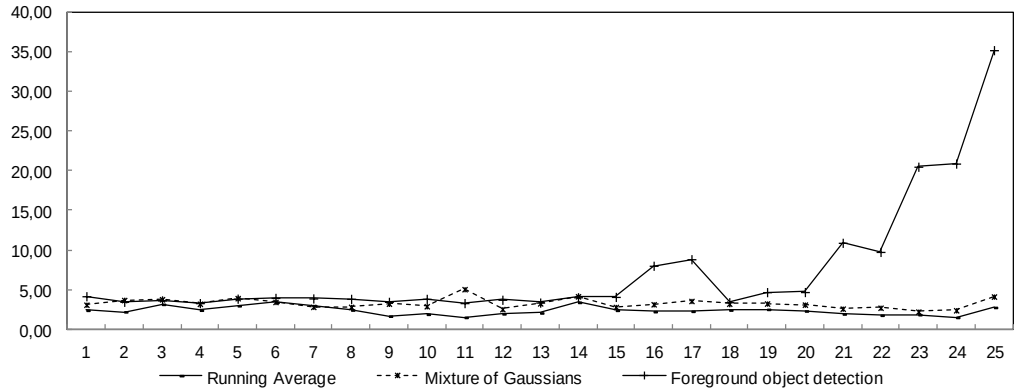


Figure 4.11 – Mean Euclidean distances obtained using the different studied segmentation models for each test image of the NADA image sequence.

The objective evaluation confirmed that the simple difference method had the worst performance with an F-measure of 69.15%. So, by analyzing the other three methods, another clear result arises; that is, the foreground object detection method fails in the last 5 test images and is generally outperformed by the other two methods. The method with better results is the running average, with the best F-measure, mean of 87.46% and standard deviation of 0.95%; however, it is worth mentioning that the mixture of Gaussians model also provides very good results, with a mean F-measure of 86.06% and standard deviation of 1.25%.

For every tested image, considering all segmentation error measures, the winning method is without a doubt the running average method. The mean Hausdorff segmentation error of contour extracted from the silhouette rounds 1.08 pixels with standard deviation of 0.07, for images with size 160x120 pixels. The Euclidean mean error, obtained from the 100 specific extracted points of the human contour is 2.35 pixels with standard deviation of 0.54. From these results, one can conclude that the running average method is capable of successfully segmenting the NADA image sequences with high quality results.

Table 4.3 – Mean and standard deviation (mean ± std) errors of the segmentations obtained using the NADA image sequence for different segmentation models.

Models	F-measure (%)	Hausdorff	Euclidean
Simple difference	69.15 ± 3.61	4.08 ± 1.16	8.52 ± 2.91
Running average	87.46 ± 0.95	1.08 ± 0.07	2.35 ± 0.54
Mixture of Gaussians	86.06 ± 1.25	1.47 ± 0.15	3.27 ± 0.63
Foreground object detection	76.00 ± 9.25	2.03 ± 0.32	7.33 ± 7.30

For the CASIA-A image sequence, 26 images were used to test the background subtraction models.

Figure 4.12 presents examples of three images of the tested images, the ground truth segmentation and the segmentation images obtained using the four background subtraction models. It is straightforward that the simple difference model cannot deal with a more complex background presenting weak segmentation results. The mixture of Gaussians model also has problems to

correctly segment the human legs, probably due to their shadows, while the other two methods seem to obtain satisfactory segmentation results.

Figure 4.13, Figure 4.14 and Figure 4.15 present the mean segmentation errors of the F-measure, Hausdorff and Euclidean distances, for each tested image for the three best models, namely the running average, mixture of Gaussians and the foreground detection models.

From Figure 4.13 it is possible to observe that the mixture of Gaussians model has lower results in the 11^{th} and 12^{th} images used in testing, and the running average has lower results in the 16^{th} to 19^{th} images. The 11^{th} test image corresponds to the central images of Figure 4.12, where mixture of Gaussians model fails to correctly segment the legs of the subject, and this failure is clearly seen in the quantitative errors. The 16^{th} to 18^{th} images and the resulting segmentation images using the running average method are presented in Figure 4.16, in which the model fails to segment the upper part of the human body. While the mixture of Gaussians blends the black band of the ground with the person, the running average has same difficulty in separating the upper body from the dark window of the background. However, it is important to mention that, once the subject passes by these obstacles, both models are capable of obtaining complete and accurate silhouettes again. On the contrary, the previously mentioned errors cannot be seen in Figure 4.14, because the Hausdorff distance does not compare the contours point by point but by the point of the ground truth contour with the closest point of the segmentation contour, which can induce small Hausdorff distances even when the contours are not so similar. The Euclidean distances, point by point, as presented in Figure 4.15, partially overcome this problem; however, it is important to remember that the extracted 100 contour points, with 45 points belonging to the left side of the subject, another 45 points belonging to the right side and 10 points are extracted from the contour between the feet of the subject. Therefore, the mean Euclidean segmentation errors of the running average model for the 16^{th} to 18^{th} images are very high because they refer to upper body contours, but the mean segmentation errors of the mixture of Gaussians models for the 11^{th} and 12^{th} images are only slightly worse than the other models, because only 10 out 100 contour points have bad results corresponding to the contour points between feet.

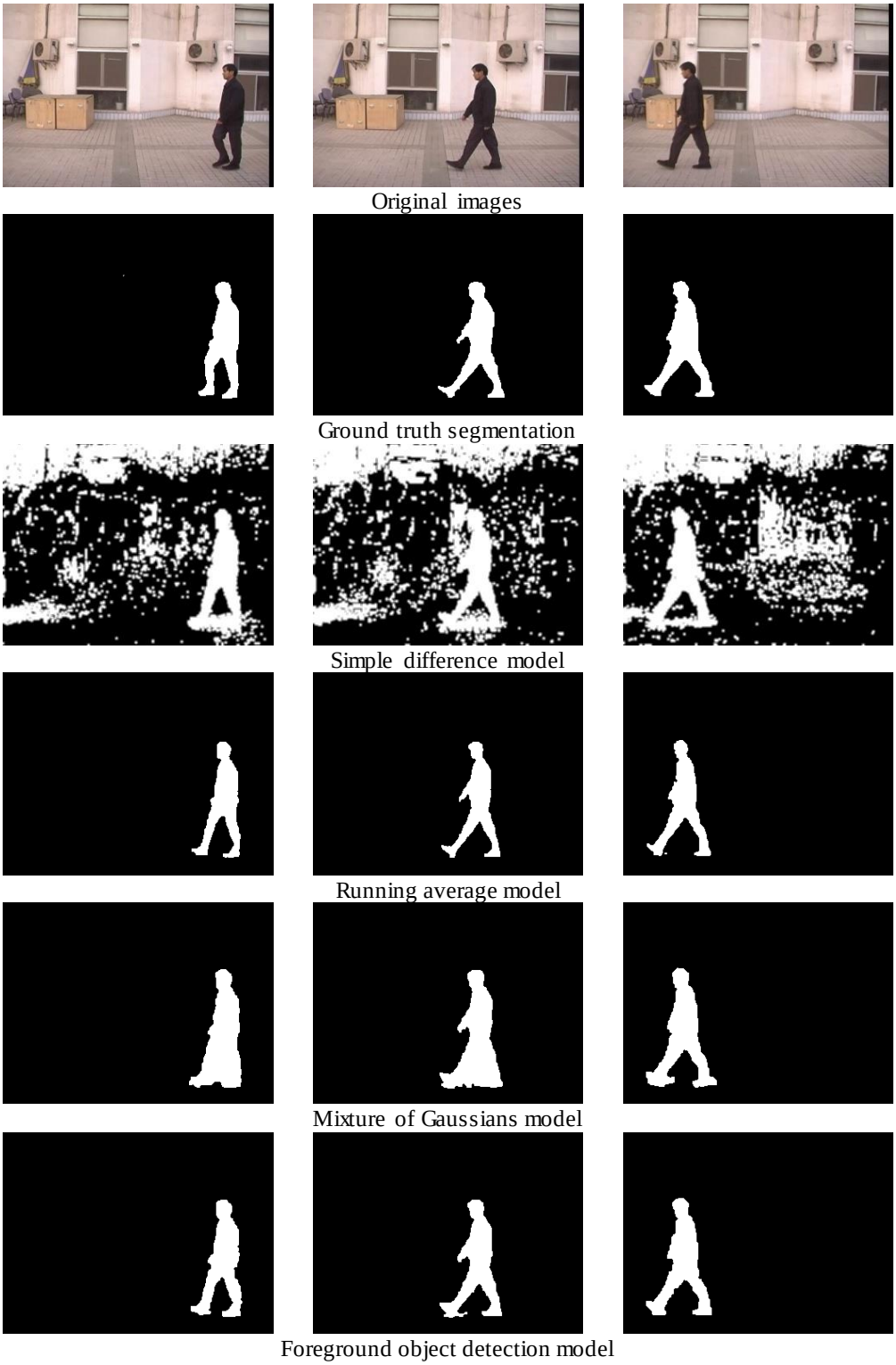


Figure 4.12 – Three images of the CASIA-A image sequence, the respective silhouette ground truth and segmentation results using the different background subtraction models.

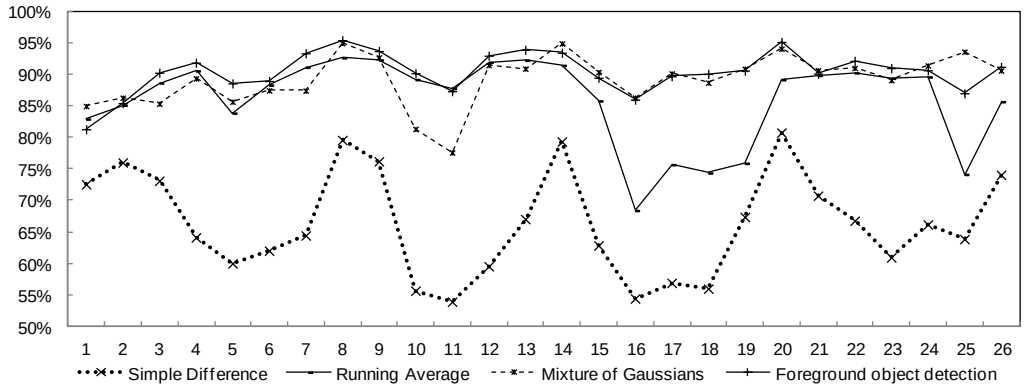


Figure 4.13 – Mean F-measures obtained using the different studied segmentation models for each test image of the CASIA-A image sequence.

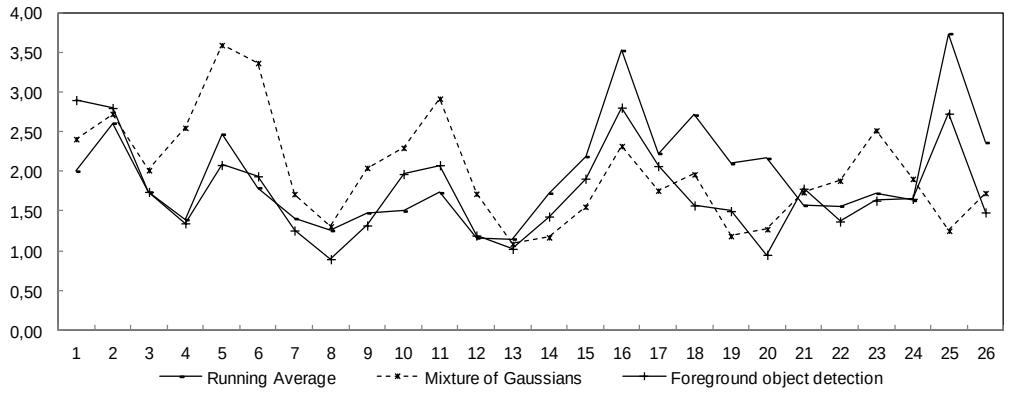


Figure 4.14 – Mean Hausdorff distances obtained using the different studied segmentation models for each test image of the CASIA-A image sequence.

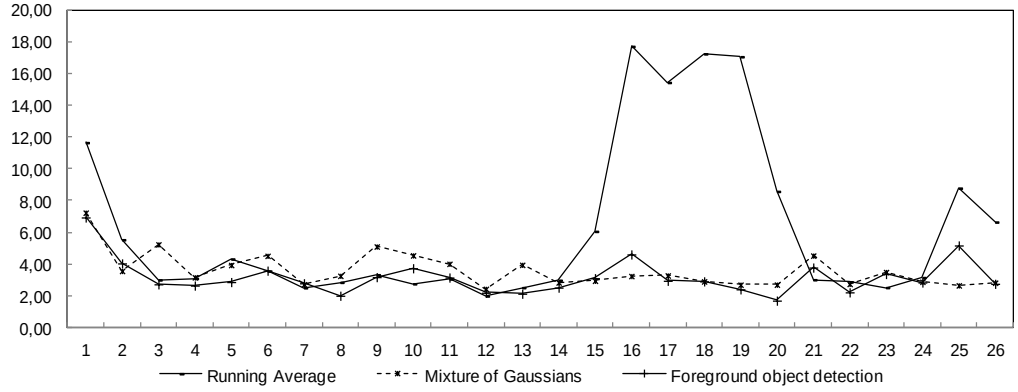


Figure 4.15 – Mean Euclidean distances obtained using the different studied segmentation models for each test image of the CASIA-A image sequence.

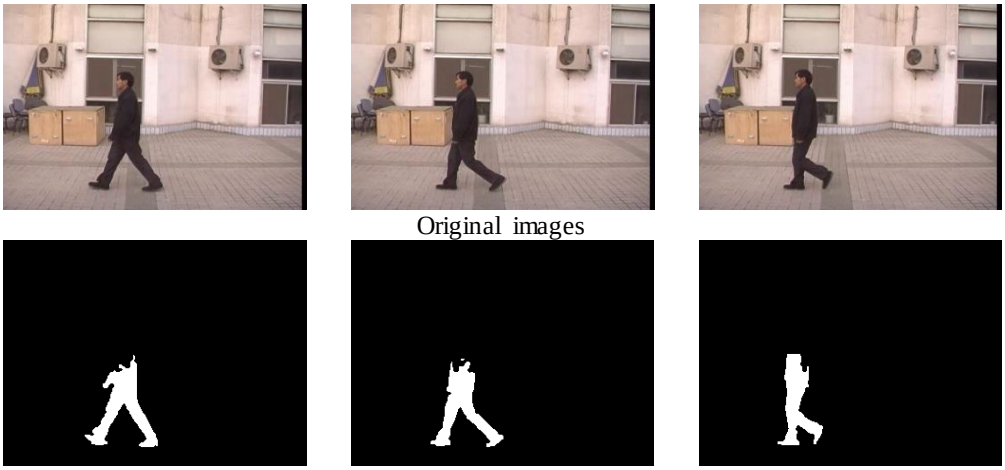


Figure 4.16 – 16th, 17th and 18th test images of the CASIA-A image sequence and the respective segmentation results using the running average model.

Table 4.4 shows the mean segmentation errors and standard deviations for the four models in the study. Since the simple difference model obtained weak results regarding the F-measure, we did not calculate the Hausdorff and Euclidean distances as it would not add useful information.

Taking into consideration all the presented results, one can conclude that the foreground object detection model provides the best results for the CASIA-A image sequence. This model is able to obtain mean Euclidean errors of 3.16 pixels with standard deviation of 1.11 pixels, from images of size 320x240 pixels.

Table 4.4 – Mean and standard deviation (mean ± std) errors of the segmentations obtained using the CASIA-A image sequence for different segmentation models.

Models	F-measure (%)	Hausdorff	Euclidean
Simple difference	66.31 ± 8.18	-	-
Running average	85.97 ± 6.74	1.96 ± 0.65	6.24 ± 5.18
Mixture of Gaussians	89.09 ± 4.05	1.99 ± 0.66	3.59 ± 1.09
Foreground object detection	90.34 ± 3.20	1.75 ± 0.57	3.16 ± 1.11

The third image sequence in study was the CAVIAR image sequence where a woman is walking through a corridor in a shopping center and the background shows the interior of a shop. A total of 22 images were used to test the segmentation results in this image sequence.

Examples from three images randomly selected of the test set are shown in Figure 4.17, together with the ground truth segmentation and the segmentation results using the four background models. The only model able to obtain complete silhouettes during most of the test images is the mixture of Gaussians model, while all the remaining models fail. Thus, we present the mean values of F-measures in Figure 4.18 for all models and the mean values of Hausdorff and Euclidean distances for the mixture of Gaussians model in Figure 4.19 and Figure 4.20, respectively. In addition, the mean and standard deviation from the mean errors of all tested images are presented in Table 4.5.

In the first test images, the four models have similar behaviors, as shown in Figure 4.18. However, when the subject approaches the center of the image, the segmentation accuracy decreases drastically in all models with the exception of the mixture of Gaussians model. In fact, the background is very complex in the center of the image and most models cannot deal with this complexity level translating into less robust segmentations where the human silhouette appears fractioned. As a result, and because the silhouette is fractioned, the final silhouette for these three models only considers the larger area, as dictated by the rule presented in section 4.2.5.

Taking into consideration all the presented results, one can conclude that the mixture of Gaussians model provides the best results for the CAVIAR image sequence. This model is able to obtain mean Euclidean errors of 6.86 pixels with standard deviation of 3.24 pixels, from images of size 320x240 pixels. The results are not as accurate as those obtained with the previous image sequences, but it should be noted that this image sequence is more complex than the others.

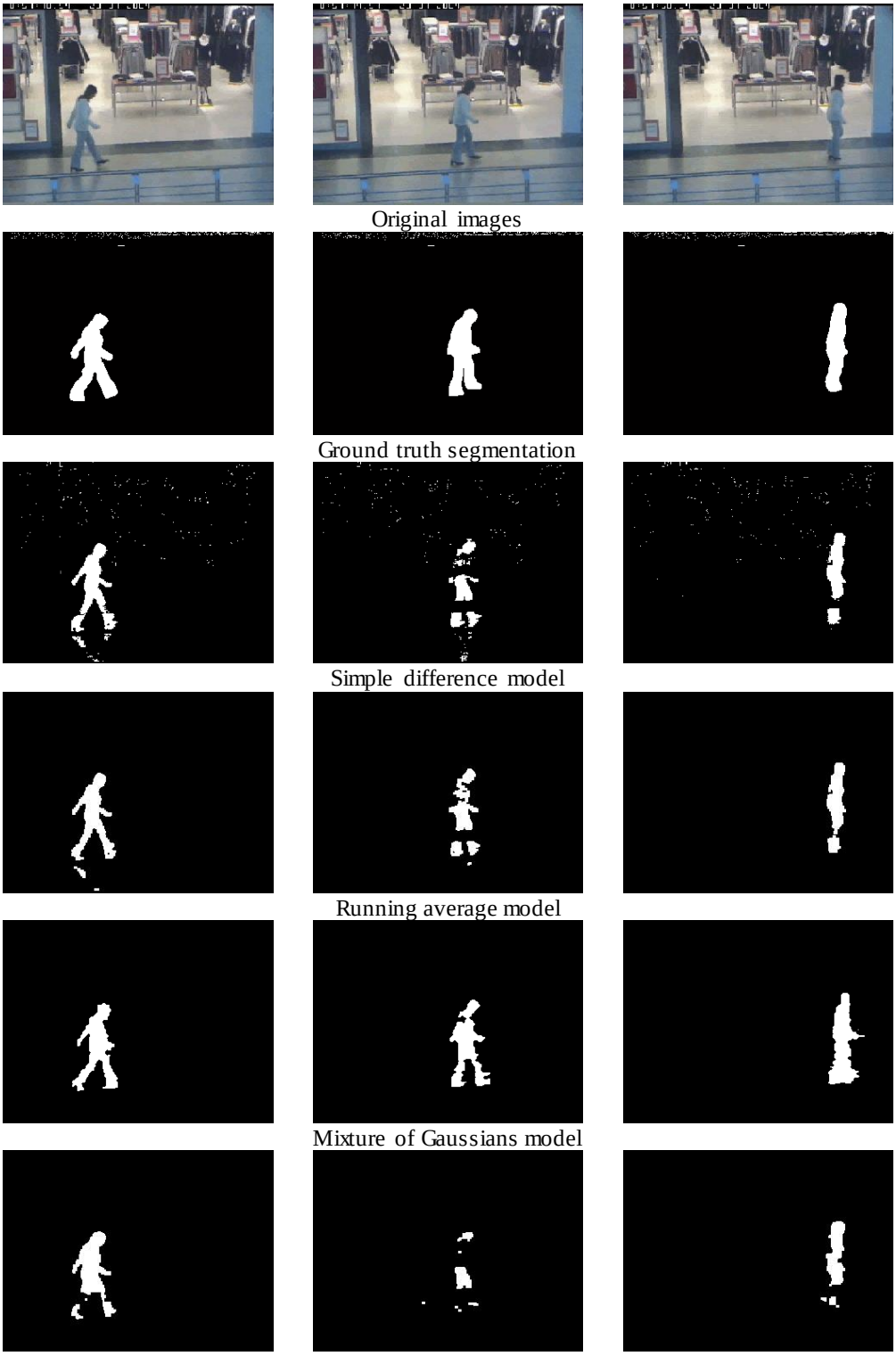


Figure 4.17 – Three images from the CAVIAR image sequence, the respective silhouette ground truth and segmentation results using the different background subtraction models.

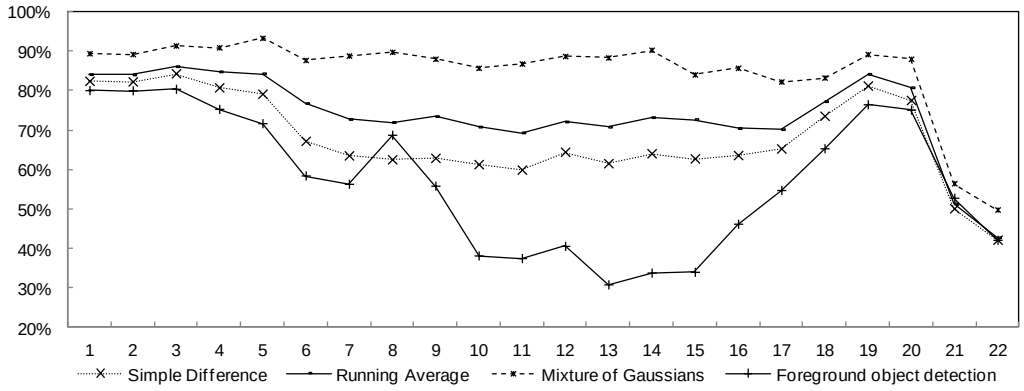


Figure 4.18 – Mean F-measure obtained using the different studied segmentation models for each test image of the CAVIAR image sequence.

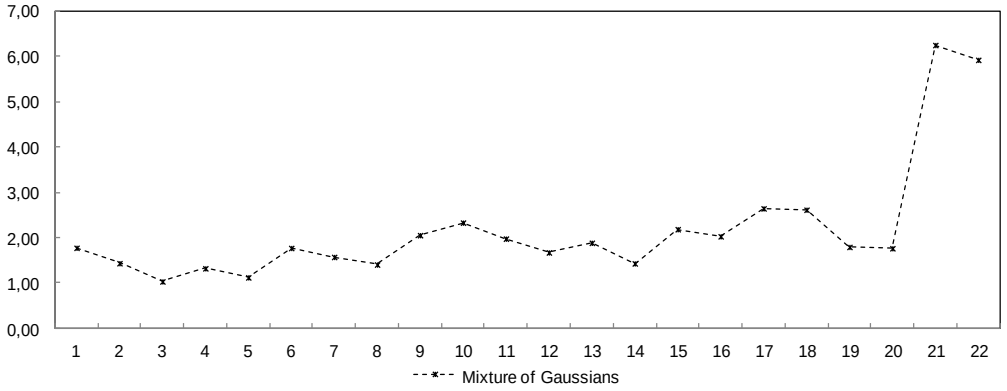


Figure 4.19 – Mean Hausdorff distances obtained using the mixture of Gaussian models for each test image of the CAVIAR image sequence.

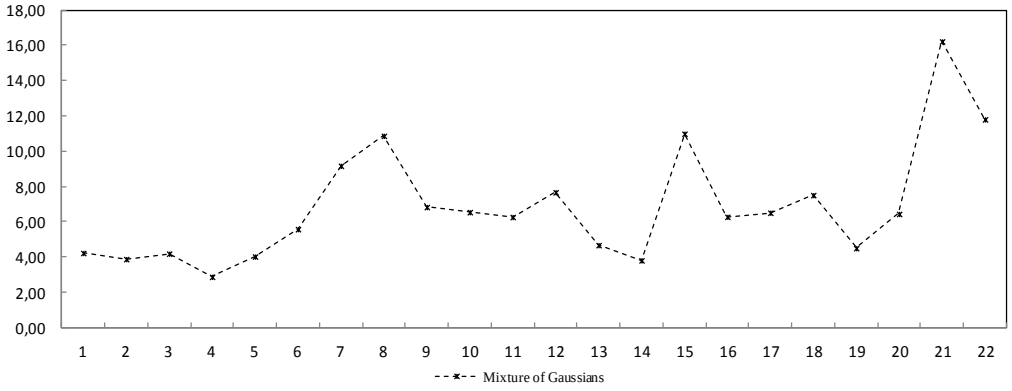


Figure 4.20 – Mean Euclidean distances obtained using the mixture of Gaussian models for each test image of the CAVIAR image sequence.

Table 4.5 – Mean and standard deviation (mean ± std) errors of the segmentations obtained using the CAVIAR image sequence for different segmentation models.

Models	F-measure (%)	Hausdorff	Euclidean
Simple difference	67.77 ± 11.05	-	-
Running average	73.76 ± 10.50	-	-
Mixture of Gaussians	84.79 ± 10.65	2.18 ± 1.33	6.86 ± 3.24
Foreground object detection	56.95 ± 17.18	-	-

The last image sequence used to test the background subtraction models was the widely used CASIA-B. Image sequences with a subject walking in a controlled environment were captured from 4 different view angles (0º, 36º, 54º and 90º). Sequences from 11 subjects were considered for this study.

Figure 4.21 shows segmentation results obtained for an image sequence with the subject walking in direction 0º. The original images and the ground truth segmentation are presented as well as the results using the four background subtraction models. From the observation of these three example images one can conclude that the simple difference model has issues in controlling the shadows. In relation to the remaining models and because the subject is walking in direction to the camera, the models tend to consider the subject as part the background.

The former problem does not occur when different directions are considered, namely 36º, 54º and 90º, as depicted in Figure 4.22. The simple difference model continues to fail, not being able to control shadows, but the remaining models, from a subjective analysis, obtain very good segmentation results.

An objective analysis is made and the results obtained for one subject are presented in Table 4.6. For the direction 0º, where the subject is walking towards the camera, the model with better segmentation results is the running average model, obtaining a mean F-measure of 89.91% and 5.01% of standard deviation with the other models having much worse results. For the other directions, the behavior of the models is similar, with simple difference models achieving the worst results, followed by the foreground object detection model. The running average model and the mixture of Gaussians obtain the best results, with the former outperforming the latter.

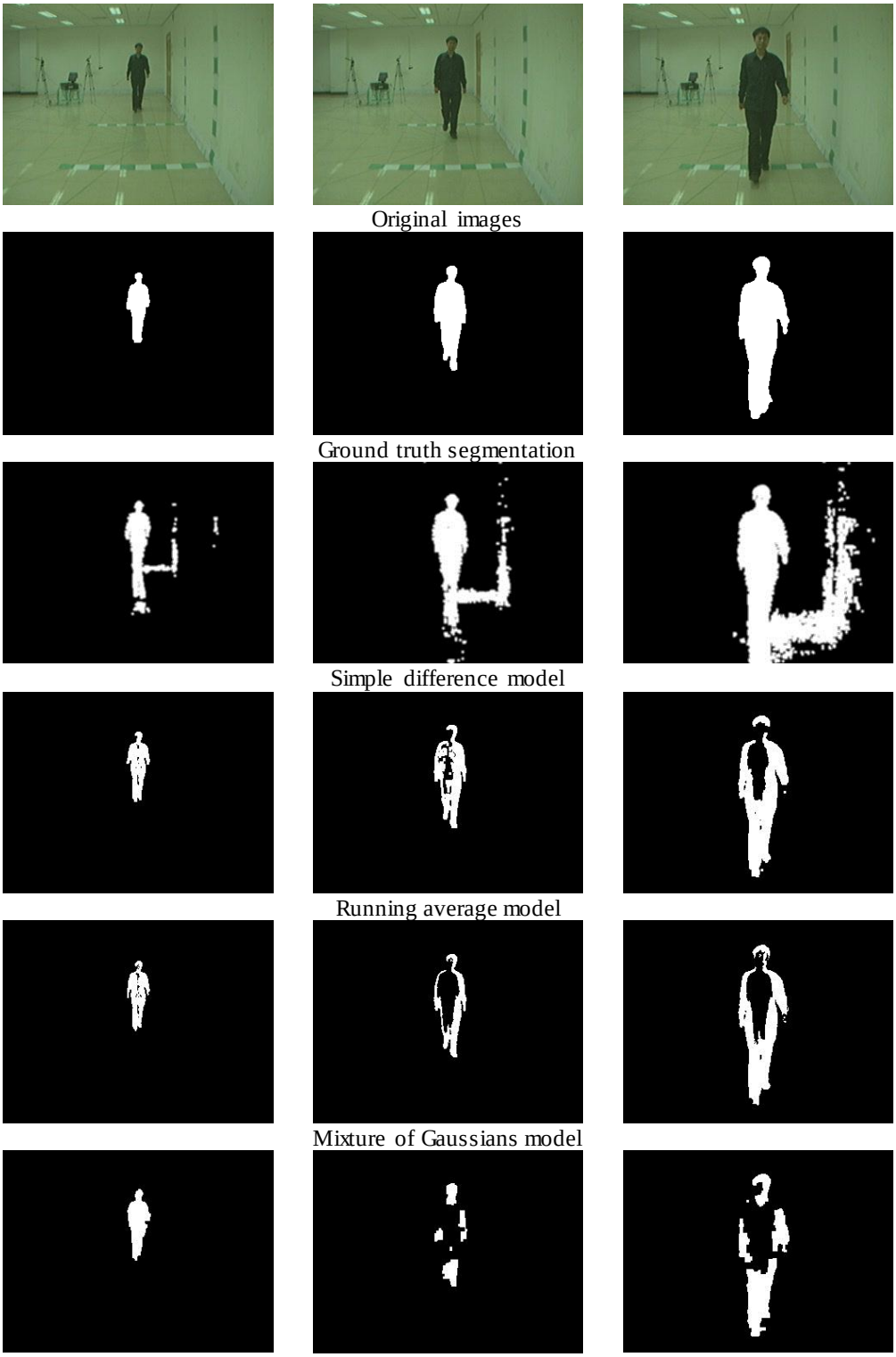


Figure 4.21 – Three images from the CASIA-B image sequences from one direction, 0º, the respective silhouette ground truth and segmentation results using the background subtraction models.

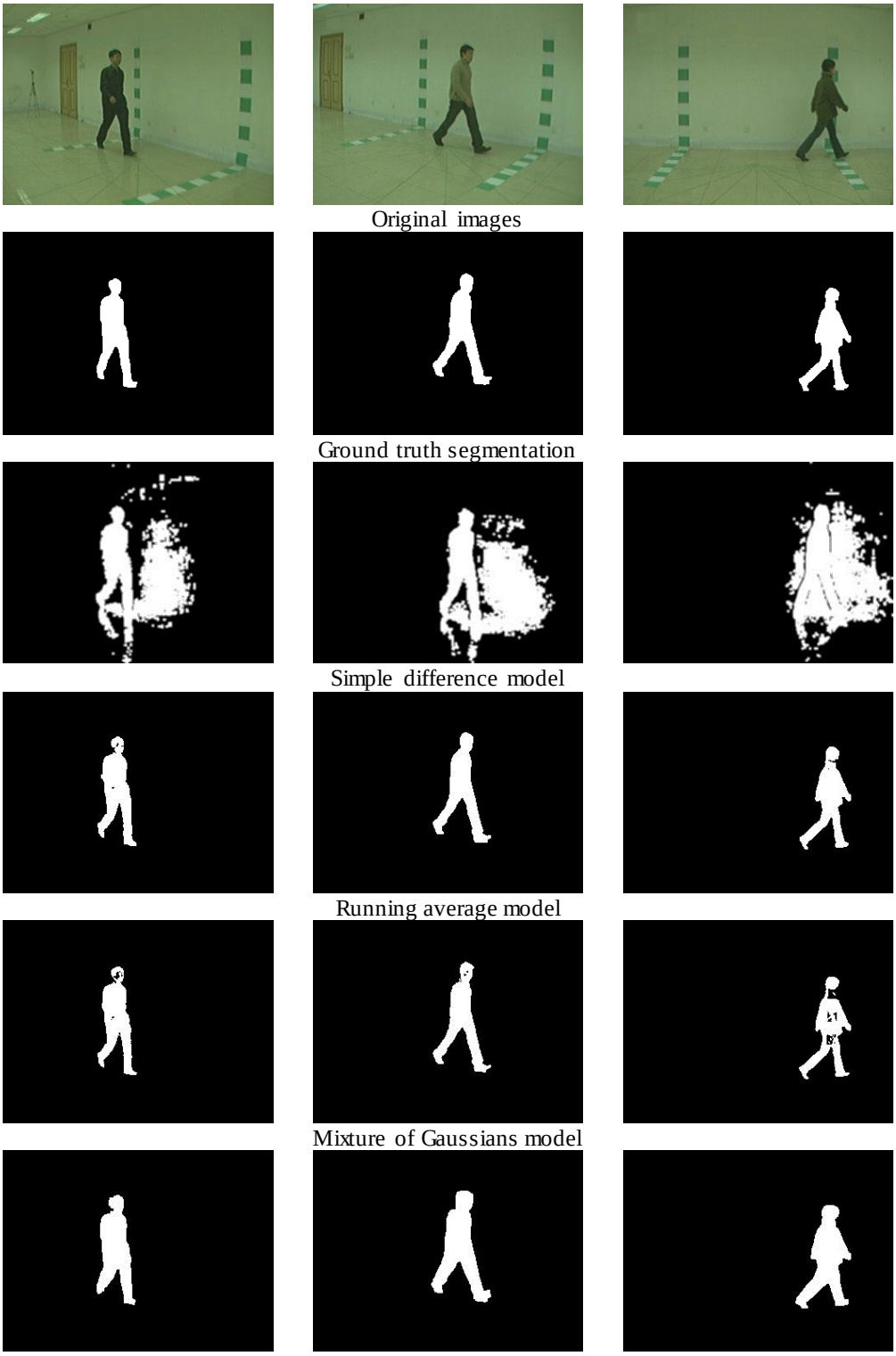


Figure 4.22 – Three images from the CASIA-B image sequences from different directions (36º, 54º and 90º), the respective silhouette ground truth and segmentation results using the background subtraction models.

Table 4.6 – Mean and standard deviations (mea n± std) of the F-measures (%) obtained using the different segmentation models for different directions studied.

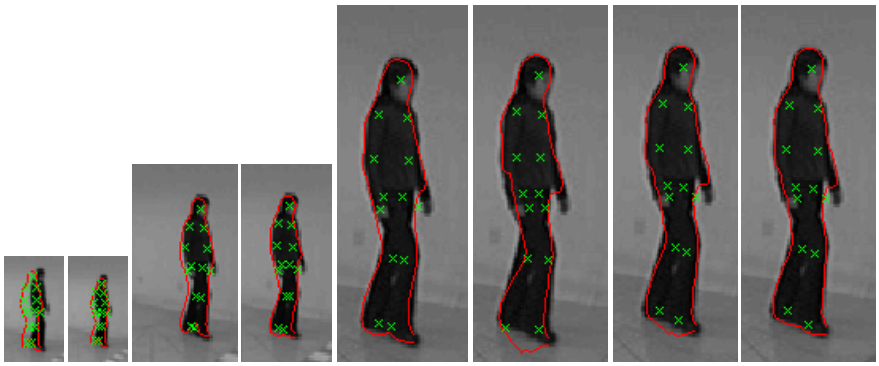
Models	0º	36º	54º	90º
Simple difference	82.15 ± 2.11	74.68 ± 5.74	67.75 ± 6.52	59.46 ± 5.85
Running average	89.91 ± 5.01	97.18 ± 1.80	98.45 ± 0.44	98.70 ± 0.29
Mixture of Gaussians	72.45±14.44	94.52 ± 2.53	96.47 ± 1.36	96.85 ± 1.11
Foreground object detection	70.75±17.73	91.50 ± 2.15	90.04 ± 1.40	88.60 ± 1.17

To sum up, the model with best segmentation results for the CASIA-B image sequences is the running average model. Although it also has problems in segmenting the silhouettes when the subject is walking towards the camera, it is able to achieve satisfactory results. In relation to the other directions studied, the running average model is capable of attaining very accurate results compared with the ground truth silhouettes.

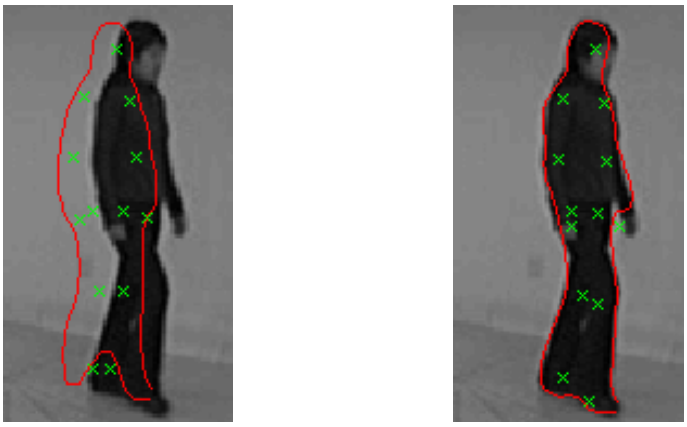
4.6.2. Active Silhouette Model

Here the segmentation results are presented, obtained using the active silhouette model described in section 4.3. The model was built using 2734 images, and 215 images were used to test the accuracy of the segmentation.

Figure 4.23 shows several steps of the active search performed in order to segment an image from the testing set, not included in the training set used. The adaptation of the ASM built throughout the iteration process to reach an optimal result can be seen in Figure 4.23. Other examples, in this case, just showing the final position, are represented in Figure 4.24. From the observation of these images, one can perceive that the landmarks corresponding to contour points have more reliable behavior than those corresponding to the anatomical points. In fact, this behavior was expected since the ASM searches for the gray level information around the point positions, and the anatomical landmark points are more likely to have similar neighbors around them, making it more difficult to choose the correct position compared with the landmark contour points.

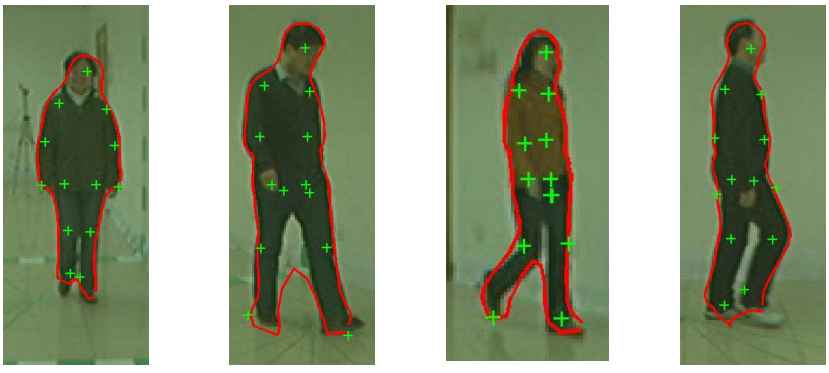


Iteration process of ASM



Initial position Final position

Figure 4.23 – Example of the iteration process using an active shape model in a new image in the first row (different image sizes correspond to different resolutions), and, in the second row, the initial and final (i.e. the segmentation result) positions of the model.



0º 36º 54º 90º

Figure 4.24 – Examples of segmentation results obtained in images for the 4 directions studied.

In order to conclude on the variation errors of the model, besides the mean Euclidean error distribution calculation for all 113 points, one also decided to study the mean Euclidean error distribution for each subgroup of points: the contour landmarks and the anatomical landmarks (i.e. from the stick figure), separately. In other words, the error distribution was calculated using 113, 100 and 13 points, corresponding to the all the points, the contour points and the stick model.

As expected, from the observation of data distribution obtained, see Figure 4.25, one can confirm that the mean error distribution is slightly worse for the subgroup of the stick points. If we take into account that the images under study have 320x240 pixels in size, it is worth noting that the results achieved with the suggested segmentation model are extremely satisfactory, within the 25^{th} to 75^{th} percentile interval ranging from 4 to 7 pixels, which translates into very accurate segmentation results. Even the mean error distribution considering only the stick points achieves good results, with the 25^{th} to 75^{th} percentile ranging from 4 to 8 pixels.

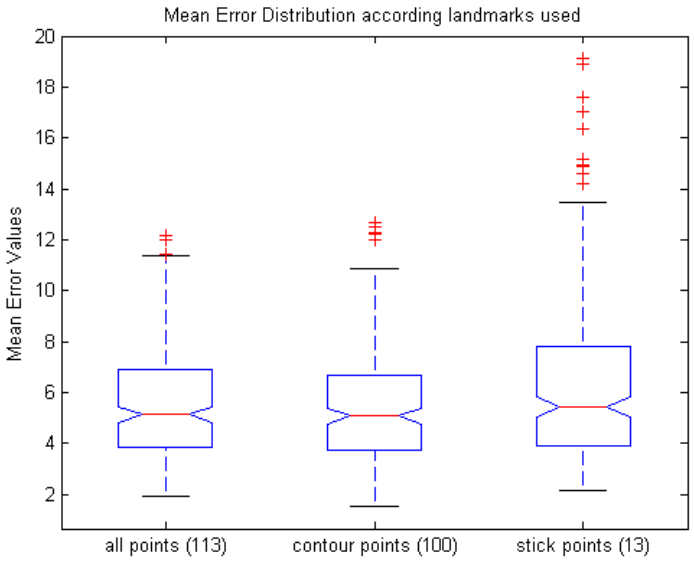


Figure 4.25 – Mean error distribution according the landmark point set (red lines are the median values and red + the outliers).

We have also studied the segmentation quality in terms of the direction in which the subjects are walking, Figure 4.26 to Figure 4.28. Analyzing these

figures, an interesting conclusion can be drawn: considering all directions against each separated direction, equivalent error ranges were achieved. This behavior is repeated independently of the landmark points.

One point it is worth noticing is in regards to error distribution of the stick figures in the 90º direction, in Figure 4.28, with the 25th to 75th percentile ranging from 8 to 13 pixels; meaning that the active silhouette model is less capable of adapting to these shapes; this result was also expected. The active silhouette model was built for the mean silhouette shape to be able to vary around one standard deviation from the mean, see Figure 4.7, meaning that the model shape is restricted. This restriction essentially translates into poor quality segmentation for images where the silhouette shapes vary more, as is the case in the 90º direction.

Table 4.7 presents a summary of the previous results, the mean and standard deviations of the mean Euclidean distances considering all landmark points or each subgroup of points and the direction of the subjects and results, once more, confirming our statements.

Briefly, it can be concluded that active silhouette models could successfully segment human shapes from images independent of the walking direction of the subjects, which is an important achievement in efficient human segmentation.

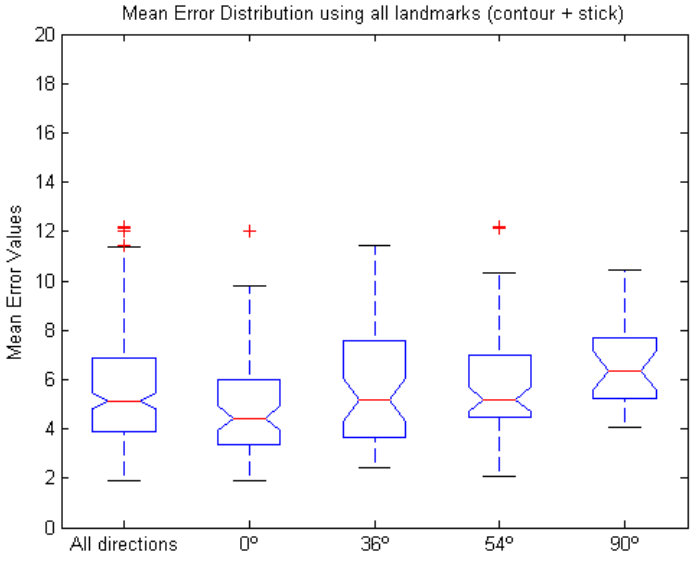


Figure 4.26 – Mean error distributions according to the direction of the subjects and considering all the 113 landmark points (red lines are the median values and red + the outliers).

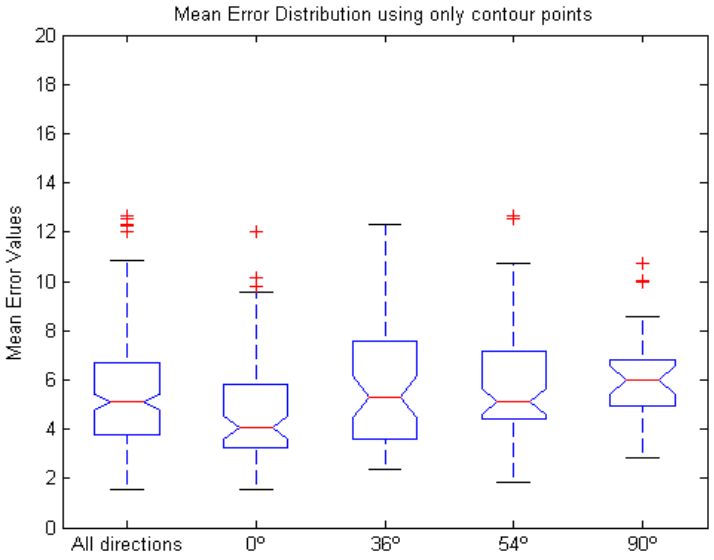


Figure 4.27 – Mean error distributions according to the direction of the subjects and considering only the landmark points from the contour (red lines are the median values and red + the outliers).

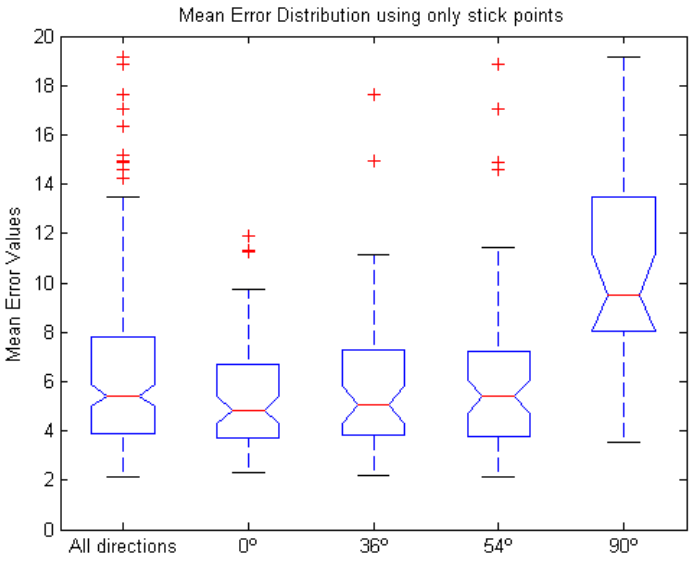


Figure 4.28 – Mean error distributions according to the direction of the subjects and considering the anatomical landmark points (red lines are the median values and red + the outliers).

Table 4.7 – Mean and standard deviation (mean ± std) errors of the mean Euclidean distributions according the direction of the subjects and the considered points.

Directions	All points (113)	Contour points (100)	Stick points (13)
All	5.58 ± 2.26	5.48 ± 2.36	6.32 ± 3.32
0º	4.84 ± 2.02	4.77 ± 2.14	5.41 ± 2.11
36º	5.77 ± 2.42	5.77 ± 2.57	5.84 ± 3.06
54º	5.85 ± 2.32	5.80 ± 2.41	6.21 ± 3.46
90º	6.62 ± 1.81	6.15 ± 2.03	10.26 ± 0.79

4.7. Discussion and Conclusion

Throughout this chapter, four background subtraction models were studied to segment humans in motion, with different levels of complexity. The first model performs differences throughout the image sequence to obtain the objects in motion; the second, more complex, applies the running average method to segment moving objects; the third, model the background with a mixture of Gaussians; and, finally, the last model uses both temporal and background differences followed by a Bayesian decision rule to classify pixels as belonging to foreground or background.

An active silhouette model was also presented in this chapter, which uses information about the contour of the silhouette together with anatomical stick points and combines the shape model with its gray level profiles for the purpose of segmenting the modeled silhouettes in new images.

Four distinct image sequences were used to apply and evaluate the abovementioned models, namely NADA, CASIA-A, CAVIAR and CASIA-B, in which the majority of sequences are part of broadly known image datasets. In the first three datasets, one subject is walking along the scene with a fixed direction of 90º in relation to the camera, while in CASIA-B the subjects are walking in four different directions. Moreover, different levels of background complexity are present throughout the datasets: the first is an outdoor scene with a homogeneous background; the second image sequence was also recorded outdoors where the background had a higher level of complexity; the third was taken inside a shopping center where an interior of a clothes shop can be seen in the background; and finally, the fourth is an indoor sequence.

With the purpose of assessing the segmentation quality of the presented methods, besides the subjective evaluation from the observation of the segmented images, three objective measures were taken into consideration throughout this work. The first measure, the F-measure, considers the equivalent rectangle that contains the ground truth silhouette image, while the second and third measures, Euclidean and Hausdorff distance, consider only the silhouette contours, or landmark points, instead.

Regarding the NADA image sequence, the running average model proved capable of successfully segmenting the human silhouettes, attaining high quality results. The mean Hausdorff segmentation error of the contour extracted from the silhouette was 1.08 pixels with standard deviation of 0.07, for images with 160x120 pixels in size. In addition, the mean Euclidean segmentation error was 2.35 pixels with standard deviation of 0.54 pixels.

About the CASIA-A image sequence, the foreground object detection model provided the best segmentation results, with the model being able to obtain mean Euclidean errors of 3.16 pixels with standard deviation of 1.11 pixels, from images of size 320x240 pixels.

As regards to CAVIAR image sequence, the mixture of Gaussians model attained the best segmentation performance. The model was able to obtain mean Euclidean errors of 6.86 pixels with standard deviation of 3.24 pixels, from images of size 320x240 pixels. The results are of inferior quality compared with the segmentation errors of the previous image sequences, but it should be noted that this image sequence is more complex than the others.

Regarding the CASIA-B image sequences, they differ from the previous ones in the sense that the subject is walking in four different directions in relation to the camera view, meaning one single direction in each image sequence. In addition, there are different subjects while in the other three datasets only a single subject is present. Concerning the results obtained for these image sequences, the model with best segmentation results was the running average model. The aforementioned background model was capable of attaining highly accurate results in three of the four directions, namely, 36º, 54º and 90º, however it had

some problems to segment the silhouettes and to achieve satisfactory results when the subject was walking towards the camera.

Our intention in using of the CASIA-B dataset was to be able to build a full model of the human silhouette that could gather the shape information of the silhouette and characterize its possible variations. Therefore, the one developed active silhouette model, which could not only model the human silhouette but also be used for segmentation purposes. In addition to the information on the silhouette, one could also infer on the position of 13 important anatomical points such as the head, shoulders, elbows, right and left hip positions, knees and feet. As a result, the active silhouette model presented here was able to achieve mean Euclidean segmentation results of 5.58 pixels with a standard deviation of 2.26 pixels, in images with 320x260 pixel size.

In conclusion, the good results obtained through the use of the active silhouette model built to perform human shape segmentation in new images strongly suggest that this type of deformable model can be used in this task. In addition, it was confirmed that just one segmentation model gathers the necessary information to segment human body structures independent of the walking direction of the subjects.

5

Conclusion and Future Work

This Thesis aimed to present computational algorithms for object segmentation and analysis in images suitable for application in objects such as the human vocal tract and the human silhouette in images.

The first challenge of this project consisted of reviewing and understanding the most appropriate and efficient methods to successfully segment objects in images. The search was directed to two objects to be modeled for the human vocal tract and silhouette, and a detailed review of the state of the art was performed towards this ends.

5.1 Application in Studying the Human Vocal Tract

The main aim for studying the vocal tract in images is to provide a better understanding of the vocal tract morphology and the movements involved in speech production. Thus, the algorithms developed to represent the vocal objects from a global perspective were based on statistical deformable models, namely active shape models and active appearance models. The primary objective consisted of the analysis of the vocal tract during the articulation of European Portuguese (EP) sounds, followed by the evaluation of the results concerning the automatic segmentation of the modeled vocal tract in new images.

The most commonly accepted imaging technique used to study the shape of the vocal tract and its articulators is magnetic resonance imaging (MRI), with the key advantages being the quality and resolution of soft-tissues and the use of non-ionizing radiation. In addition, and taking into consideration that an efficient deformable model is strictly related to the quality of the images, two datasets with different image qualities were analyzed in this Thesis, one acquired using a 1.5 Tesla (1.5T) MRI system and another using a 3.0 Tesla (3.0T) MRI system, with the latter having higher resolution.

From the experimental results obtained, one may conclude that the statistical deformable models built are capable of efficiently characterizing the behavior of the vocal tract modeled from MR images. Moreover, the modes of variation of the models constructed were able to provide further explanation of the actual actions involved in the EP speech sounds considered.

While active shape models consider the information around each landmark point of the modeled object, active appearance models also use the gray level information of the object under study. Consequently, the former types of models tend to be less efficient than the latter, which was consistent with our findings in this Thesis. Nevertheless, both models obtained remarkable results, either in terms of translating the movements and configurations involved in speech production, as well as in the segmentation and characterization of the vocal tract in new images.

Regarding the segmentation results with respect to image quality, for the 1.5T dataset, where 256 x 256 pixels images were used, mean errors rounding 10 pixels were achieved; while the segmentation results using the 3.0T MR images led to similar mean errors but in double-sized images, 512 x 512 pixels. Thus, the results obtained confirm that segmentation is more accurate when images with better quality are used, such as the dataset acquired with 3.0T MRI.

Another contribution accomplished in this Thesis concerns the amount of data used to characterize and segment the vocal tract during speech production. For the 3.0T sounds dataset, 25 out of 30 possible EP speech sounds were modeled for two subjects, male and female, in addition to using three measurements (slices) for each sound, translating in a dataset with a total of 150

images, in which 138 images were used for training the model and 12 images used for testing.

To conclude, the usage of active shape models and active appearance models made possible the automatic and realistic simulation of the vocal tract during speech production as well as the efficient segmentation and characterization of vocal tract in new images. Furthermore, the use of such automatic image analysis techniques can allow for obtaining quantitative measures with higher precision, being particularly advantageous when a large volume of data must be analyzed.

5.2 Application on Human Silhouette

In this Thesis, four background subtraction models were studied to segment human silhouettes in image sequences, with different levels of complexity. The simplest model calculates differences throughout the image sequence to obtain the objects in motion; the second, being more complex, applies the running average method; the third, models the background with a mixture of Gaussians; and, finally, the last model uses both temporal and background differences followed by a Bayes decision rule to classify pixels as belonging to foreground or background.

In addition to the background subtraction models, and following the methodology used for modeling the vocal tract, an active silhouette model was also developed, using information about the contour of the silhouette together with anatomical stick points and combining the shape model with its gray level profiles with the aim of segmenting the modeled silhouettes in new images.

Four distinct image sequences were used to built and evaluate the abovementioned models, namely NADA, CASIA-A, CAVIAR and CASIA-B, where the majority of sequences are part of broadly known datasets. In all datasets only one subject is present in the images, whereas in the first three the subject walks in one fixed direction while in the last dataset the subject walks in four

different directions in the sequences. Furthermore, different levels of background complexity are present throughout the datasets.

Both subjective and quantitative assessment was used for the obtained segmentation results. Regarding the NADA image sequence, the running average model attained the best segmentation results compared with the other models. As to the CASIA-A image sequence, the background subtraction model that obtained better results was the foreground object detection model. As regards to CAVIAR image sequence, the mixture of Gaussians model achieved the best segmentation performance. Concerning the CASIA-B image sequences, the model with best segmentation results was the running average model. This background model was capable of obtaining very accurate results in three of the four directions, namely, 36º, 54º and 90º; however, it had some issues in segmenting the silhouettes when the subject was walking towards the camera, achieving only satisfactory results.

Besides testing the segmentation results obtained by background subtraction models, an active silhouette model was developed that was not only capable of modeling the human silhouette but was also able to be used for segmentation purposes. A particular feature of this model is the possibility of inferring the position of 13 important anatomical points such as the head, shoulders, elbows, right and left hip positions, knees and feet, in addition to information on the silhouette.

Since different background subtraction achieved best segmentation results, no consensus exists on choosing the best segmentation method based on background subtraction for segmenting silhouettes in images. The choice greatly depends on the level of complexity of the images. On the contrary, the excellent results obtained through the use of the active silhouette model built to perform human shape segmentation in new images strongly suggest that this type of deformable models can be used in this task.

To conclude, the main contribution regarding the segmentation of human silhouettes in images was that it allowed for building an active shape model that gathers the necessary information independent of the walking direction of the subject.

5.3 Future Work

Although the developed algorithms for the human vocal tract and human silhouette have achieved successful results, there are some aspects that still need to be improved.

Regarding the models built for the human vocal tract, it would be important to have more images of each sound from each subject as well as more images from different subjects in order to evaluate with more precision for instance the intra-subject and inter-subject variability. It would be also interesting to compare the results with other methodologies. With regards to the improvement of the vocal tract active models, the active search initialization for segmentation purposes should be refined with the development of an automatic algorithm. Another natural step would be the study and development of 3D models of the vocal tract using the 3D deformable models in order to have more realistic models.

On the subject of the human silhouette modeling and segmentation it would be important to develop methodologies that can combine, more accurately, the human silhouette shape with important anatomical joint positions, mainly for use in biomechanical studies related to human motion. As soon as the use of simple image cameras allows the robust and detailed analysis of real movements performed by subjects in their daily life, it will be possible to obtain new levels of information of the subjects from the input images. Up until now, this has been only achieved from images acquired under well controlled conditions and in significantly restricted environments, which consequently demands more robust techniques of image segmentation, motion tracking and analysis.

Bibliography

Aggarwal, J.K., and Q. Cai. 1999. "Human Motion Analysis: A Review." *Computer Vision and Image Understanding* 73 (3): 428–40. doi:10.1006/cviu.1998.0744.

Aggarwal, J.K., Q. Cai, W. Liao, and B. Sabata. 1994. "Articulated and Elastic Non-Rigid Motion: A Review." In , *Proceedings of the 1994 IEEE Workshop on Motion of Non-Rigid and Articulated Objects, 1994*, 2–14. doi:10.1109/MNRAO.1994.346261.

Aggarwal, J.K., and M.S. Ryoo. 2011. "Human Activity Analysis: A Review." *ACM Computing Surveys (CSUR)* 43 (3): 16:1–16:43. doi:10.1145/1922649.1922653.

Agin, G.J. 1980. "Computer Vision Systems for Industrial Inspection and Assembly." *Computer* 13 (5): 11–20. doi:10.1109/MC.1980.1653613.

Ahmed, R., G.C. Karmakar, and L.S. Dooley. 2006. "Probabilistic Spatio-Temporal Video Object Segmentation Incorporating Shape Information." In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP*, 2:II–II. doi:10.1109/ICASSP.2006.1660425.

Amer, A. 2003. "Memory-Based Spatio-Temporal Real-Time Object Segmentation for Video Surveillance." In *Electronic Imaging 2003. International Society for Optics and Photonics, 2003*, 5012:10–21. doi:10.1117/12.477500.

Apostol, L., P. Perrier, M. Raybaudi, and C. Segebarth. 1999. "3D Geometry of the Vocal Tract and Inter-Speaker Variability." In *Proceedings of the 14th International Congress of Phonetic Sciences*, 443–46. San Francisco, USA.

Arantes, M., and A. Gonzaga. 2011. "Human Gait Recognition Using Extraction and Fusion of Global Motion Features." *Multimedia Tools and Applications* 55 (3): 655–75. doi:10.1007/s11042-010-0587-y.

Arnal, A., P. Badin, G. Brock, and P. Connan. 2000. "Une Base de Données Cinéradiographiques Du Français." *XXIIIèmes Journées d'Etude Sur La Parole*, 425–28.

Avila-Garcia, M. S., J. N. Carter, and R. I. Damper. 2004. "Extracting Tongue Shape Dynamics from Magnetic Resonance Image Sequences." In *Proceedings of the International Conference on Signal Processing (ICSP 2004)*, 288–91.

Ayache, N. 1998. "Medical Image Analysis a Challenge for Computer Vision Research." In *Proceedings of the Fourteenth International Conference on Pattern Recognition, 1998.*, 2:1255–1256 vol.2. doi:10.1109/ICPR.1998.711928.

Badin, P., G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux. 2002. "Three-Dimensional Linear Articulatory Modeling of Tongue, Lips and

Face, Based on MRI and Video Images." *Journal of Phonetics* 30 (3): 533–53. doi:10.1006/jpho.2002.0166.

Badin, P., and A. Serrurier. 2006. "Three-dimensional linear modeling of tongue: Articulatory data and models." In *Proceedings of the 7th International Seminar on Speech Production, ISSP7*, 395–402. Ubatuba, SP, Brazil: UFMG, Belo Horizonte, Brazil.

Baer, T., J. C. Gore, S. Boyce, and P. W. Nye. 1987. "Application of MRI to the Analysis of Speech Production." *Magnetic Resonance Imaging* 5 (1): 1–7. doi:10.1016/0730-725X(87)90477-2.

Baer, T., J. C. Gore, L. C. Gracco, and P. W. Nye. 1991. "Analysis of Vocal Tract Shape and Dimensions Using Magnetic Resonance Imaging: Vowels." *The Journal of the Acoustical Society of America* 90 (2 Pt 1): 799–828.

Bakhshaee, H., C. Moro, K. Kost, and L. Mongeau. 2013. "Three-Dimensional Reconstruction of Human Vocal Folds and Standard Laryngeal Cartilages Using Computed Tomography Scan Data." *Journal of Voice* 27 (6): 769–77. doi:10.1016/j.jvoice.2013.06.003.

Ball, M.J., and O. Lowry. 2008. *Methods in Clinical Phonetics*. John Wiley & Sons.

Ballard, D.H., and C.M.B. Brown. 1982. *Computer Vision*. Englewood Cliffs, NJ: Prenice-Hall.

Barbu, T. 2014. "Pedestrian Detection and Tracking Using Temporal Differencing and HOG Features." *Computers & Electrical Engineering* 40 (4): 1072–79. doi:10.1016/j.compeleceng.2013.12.004.

Beautemps, D., P. Badin, and R. Laboissière. 1995. "Deriving Vocal-Tract Area Functions from Midsagittal Profiles and Formant Frequencies: A New Model for Vowels and Fricative Consonants Based on Experimental Data." *Speech Communication* 16 (1): 27–47. doi:10.1016/0167-6393(94)00045-C.

Beetz, M., B. Kirchlechner, and M. Lames. 2005. "Computerized Real-Time Analysis of Football Games." *IEEE Pervasive Computing* 4 (3): 33–39. doi:10.1109/MPRV.2005.53.

Begg, R.K., M. Palaniswami, and B. Owen. 2005. "Support Vector Machines for Automated Gait Classification." *IEEE Transactions on Biomedical Engineering* 52 (5): 828–38. doi:10.1109/TBME.2005.845241.

Behrends, J., P. Hoole, G. Leinsinger, H.G. Tillmann, K. Hahn, M. Reiser, and A. Wismüller. 2003. "A Segmentation and Analysis Method for MRI Data of the Human Vocal Tract." In *Bildverarbeitung Für Die Medizin 2003*, 186–90. Informatik Aktuell. Springer Berlin Heidelberg.

Benesty, J., M.M. Sondhi, and Y. Huang. 2008. *Springer Handbook of Speech Processing*. Springer Heidelberg.

Birkholz, P., and B. Kroger. 2006. "Vocal Tract Model Adaptation Using Magnetic Resonance Imaging." In *Proceedings of the 7th International Seminar on Speech Production*, 493–500.

Blake, A., and M. Isard. 1998. *Active Contours: The Application of Techniques from Graphics,Vision,Control Theory and Statistics to Visual Tracking of Shapes in Motion*. Springer-Verlag New York, Inc.

Bresch, E., Y. Kim, K. Nayak, D. Byrd, and S. Narayanan. 2008. "Seeing Speech: Capturing Vocal Tract Shaping Using Real-Time Magnetic Resonance Imaging [Exploratory DSP]." *IEEE Signal Processing Magazine* 25 (3): 123–32. doi:10.1109/MSP.2008.918034.

Bresch, E., J. Nielsen, K. Nayak, and S. Narayanan. 2006. "Synchronized and Noise-Robust Audio Recordings during Realtime Magnetic Resonance Imaging Scans." *The Journal of the Acoustical Society of America* 120 (4): 1791–94.

Campos, M., M. Ferreira, T. Martins, and C. Santos. 2010. "Inspection of Bottles Crates in the Beer Industry through Computer Vision." In *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, 1138–43. doi:10.1109/IECON.2010.5675529.

Carbone, I., P. Martins, A. Teixeira, and A. Silva. 2008. "A Vocal Tract Segmentation and Analysis over a European Portuguese MRI Database." *Electrónica E Telecomunicações* 4 (9): 1050–53.

Carvalho, F. J. S., and J. M. R. S. Tavares. 2005. "Metodologias Para Identificação de Faces Em Imagens: Introdução E Exemplos de Resultados." In *Congreso de Métodos Numéricos En Ingenieria*, 14 pages. Granada, Spain.

Carvalho, F. J. S., and J. M. R. S. Tavares. 2007. "Eye Detection Using a Deformable Template in Static Images." In *VIPimage - I ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing*, 209–15. Porto, Portugal: Taylor & Francis.

Caselles, V., R. Kimmel, and G. Sapiro. 1997. "Geodesic Active Contours." *International Journal of Computer Vision* 22 (1): 61–79.

Cedras, C., and M. Shah. 1995. "Motion-Based Recognition: A Survey." *Image and Vision Computing* 13: 129–55.

Chaaraoui, A.A., P. Climent-Pérez, and F. Flórez-Revuelta. 2012. "A Review on Vision Techniques Applied to Human Behaviour Analysis for Ambient-Assisted Living." *Expert Systems with Applications* 39 (12): 10873–88. doi:10.1016/j.eswa.2012.03.005.

Chan, T.F., and L.A Vese. 2001. "Active Contours without Edges." *IEEE Transactions on Image Processing* 10 (2): 266–77. doi:10.1109/83.902291.

Chaquet, J.M., E.J. Carmona, and A. Fernández-Caballero. 2013. "A Survey of Video Datasets for Human Action and Activity Recognition." *Computer Vision and Image Understanding* 117 (6): 633–59. doi:10.1016/j.cviu.2013.01.013.

Chen, L., H. Wei, and J. Ferryman. 2013. "A Survey of Human Motion Analysis Using Depth Imagery." *Pattern Recognition Letters*, Smart Approaches for Human Action Recognition, 34 (15): 1995–2006. doi:10.1016/j.patrec.2013.02.006.

Cheung, K., S. Baker, and T. Kanade. 2005. "Shape-From-Silhouette Across Time Part II: Applications to Human Modeling and Markerless Motion Tracking." *International Journal of Computer Vision* 63 (3): 225–45. doi:10.1007/s11263-005-6879-4.

Clément, P., S. Hans, D.M. Hartl, S. Maeda, J. Vaissière, and D. Brasnu. 2007. "Vocal Tract Area Function for Vowels Using Three-Dimensional Magnetic Resonance Imaging. A Preliminary Study." *Journal of Voice* 21 (5): 522–30. doi:10.1016/j.jvoice.2006.01.005.

CMU Graphics Lab. 2001. "CMU Graphics Lab Motion Capture Database." *CMU Graphics Lab Motion Capture Database*. http://mocap.cs.cmu.edu/ [Acessed on September 2014].

Comaniciu, D., V. Ramesh, and P. Meer. 2000. "Real-Time Tracking of Non-Rigid Objects Using Mean Shift." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2000*, 2:142–49.

Cootes, T. F. 2004. "Modelling and Search Software." http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/software/am_tools_doc/download_win.html.

Cootes, T. F., G. J. Edwards, and C. J. Taylor. 1998. "Active Appearance Models." In *Computer Vision — ECCV'98*, 484–98. Lecture Notes in Computer Science 1407. Springer Berlin Heidelberg.

Cootes, T. F., and C. J. Taylor. 1992. "Active Shape Models - 'Smart Snakes.'" In *BMVC92*, 266–75. Springer London.

Cootes, T. F., C. J. Taylor, D. H. Cooper, and J. Graham. 1992. "Training Models of Shape from Sets of Examples." In *Proceedings of the British Machine Vision Conference*, 9–18. Leeds: Springer London.

Cootes, T. F., C. J. Taylor, D. H. Cooper, and J. Graham. 1995. "Active Shape Models-Their Training and Application." *Computer Vision and Image Understanding* 61 (1): 38–59. doi:10.1006/cviu.1995.1004.

Cootes, T.F., C.J. Taylor, and A. Lanitis. 1994. "Multi-Resolution Search with Active Shape Models." In *, Proceedings of the 12th IAPR International Conference on Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision Amp; Image Processing*, 1:610–612 vol.1. doi:10.1109/ICPR.1994.576375.

Corazza, Stefano, Lars Mündermann, Emiliano Gambaretto, Giancarlo Ferrigno, and Thomas P. Andriacchi. 2010. "Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Specific Model Generation." *International Journal of Computer Vision* 87 (1-2): 156–69. doi:10.1007/s11263-009-0284-3.

Crary, M. A., I. M. Kotzur, J. Gauger, M. Gorham, and S. Burton. 1996. "Dynamic Magnetic Resonance Imaging in the Study of Vocal Tract Configuration." *Journal of Voice* 10 (4): 378–88. doi:10.1016/S0892-1997(96)80030-0.

Cremers, D. 2006. "Dynamical Statistical Shape Priors for Level Set-Based Tracking." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (8): 1262–73. doi:10.1109/TPAMI.2006.161.

Cremers, D., M. Rousson, and R. Deriche. 2007. "A Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion and Shape." *International Journal of Computer Vision* 72 (2): 195–215. doi:10.1007/s11263-006-8711-1.

Cucchiara, R., C. Grana, A. Prati, and R. Vezzani. 2005. "Probabilistic Posture Classification for Human-Behavior Analysis." *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 35 (1): 42–54. doi:10.1109/TSMCA.2004.838501.

Damas, S., O. Cordón, and J. Santamaría. 2011. "Medical Image Registration Using Evolutionary Computation: An Experimental Survey." *IEEE Computational Intelligence Magazine* 6 (4): 26–42. doi:10.1109/MCI.2011.942582.

Dang, J., and K. Honda. 2002. "Estimation of Vocal Tract Shapes from Speech Sounds with a Physiological Articulatory Model." *Journal of Phonetics* 30 (3): 511–32. doi:10.1006/jpho.2002.0167.

Davis III, R.B., S. Õunpuu, D. Tyburski, and James R. Gage. 1991. "A Gait Analysis Data Collection and Reduction Technique." *Human Movement Science* 10 (5): 575–87. doi:10.1016/0167-9457(91)90046-Z.

Dimitrijevic, M., V. Lepetit, and P. Fua. 2006. "Human Body Pose Detection Using Bayesian Spatio-Temporal Templates." *Computer Vision and Image Understanding*, Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour, 104 (2–3): 127–39. doi:10.1016/j.cviu.2006.07.007.

Dou, Y., and J. Xu. 2007. "FPGA-Accelerated Active Shape Model for Real-Time People Tracking." In *Advances in Computer Systems Architecture*, 268–79. Lecture Notes in Computer Science 4697. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-540-74309-5_26.

Ekinci, M., and E. Gedikli. 2005. "Silhouette Based Human Motion Detection and Analysis for Real-Time Automated Video Surveillance." *Turkish Journal of Electrical Engineering and Computer Sciences* 13 (2): 199–229.

Elgammal, A., R. Duraiswami, and L.S. Davis. 2003. "Efficient Kernel Density Estimation Using the Fast Gauss Transform with Applications to Color Modeling and Tracking." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (11): 1499–1504. doi:10.1109/TPAMI.2003.1240123.

Engwall, O. 2000a. "Are Static MRI Measurements Representative of Dynamic Speech? Results from a Comparative Study Using MRI, EPG and EMA." In *INTERSPEECH*, I:17–20. Beijing, China.

Engwall, O. 2000b. "A 3D Tongue Model Based on MRI Data." In *Proceedings Sixth International Conference on Spoken Language Processing*, 901–4.

Engwall, O. 2003. "A Revisit to the Application of MRI to the Analysis of Speech Production-Testing Our Assumptions." *Proceedings of 6th International Seminar on Speech Production*, 43–48.

Engwall, O., and P. Badin. 2000. "An MRI Study of Swedish Fricatives: Coarticulatory Effects." In *Proceedins 5 Th Speech Production Seminar*, 297–300.

Fahmy, M.M. 1994. "Automatic Number-Plate Recognition: Neural Network Approach." In *Vehicle Navigation and Information Systems Conference, 1994. Proceedings., 1994*, 99–101. doi:10.1109/VNIS.1994.396858.

Fisher, R.B, J. Santos-victor, and J. Crowley. 2003. "CAVIAR: Context Aware Vision Using Image-Based Active Recognition." *EC's Information Society Technology's Programme Project IST 2001 37540*. http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/.

Fitzpatrick, L., and A. Chasaide. 2002. "Estimating Lingual Constriction Location in High Vowels: A Comparison of EMA- and EPG-Based Measures." *Journal of Phonetics* 30 (3): 397–415. doi:10.1006/jpho.2001.0158.

Fontecave, J., and F. Berthommier. 2005. "Quasi-Automatic Extraction Method of Tongue Movement from a Large Existing Speech Cineradiographic Database." In *Proceedings of the Interspeech*, 1081–84. Lisbon.

Fontecave Jallon, J., and F. Berthommier. 2009. "A Semi-Automatic Method for Extracting Vocal Tract Movements from X-Ray Films." *Speech Communication* 51 (2): 97–115. doi:10.1016/j.specom.2008.06.005.

Freifeld, O., A. Weiss, S. Zuffi, and M.J. Black. 2010. "Contour People: A Parameterized Model of 2D Articulated Human Shape CVPR." In *IEEE Conf. on Computer Vision and Pattern Recognition*. Vol. 639–46.

Fujimura, O., S. Kiritani, and H. Ishida. 1973. "Computer Controlled Radiography for Observation of Movements of Articulatory and Other Human Organs." *Computers in Biology and Medicine* 3 (4): 371–84. doi:10.1016/0010-4825(73)90003-6.

Garibotto, G. 2009. "Video Surveillance and Biometric Technology Applications." In *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009. AVSS '09*, 288–288. doi:10.1109/AVSS.2009.111.

Gavrila, D. M. 1999. "The Visual Analysis of Human Movement: A Survey." *Computer Vision and Image Understanding* 73 (1): 82–98. doi:10.1006/cviu.1998.0716.

Goldenberg, R., R. Kimmel, E. Rivlin, and M. Rudzsky. 2001. "Fast Geodesic Active Contours." *IEEE Transactions on Image Processing* 10 (10): 1467–75. doi:10.1109/83.951533.

Gonçalves, P.C.T., J.M.R.S. Tavares, and R.M.N. Jorge. 2009. "Segmentation and Simulation of Objects Represented in Images Using Physical Principles." *ICCES*, Tech Science Press, 9 (3): 203–4. doi:10.3970/icces.2009.009.203.

Gonzalez, J.J., I.S. Lim, P. Fua, and D. Thalmann. 2003. "Robust Tracking and Segmentation of Human Motion in an Image Sequence." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP '03)*, 3:III–29–32 vol.3. doi:10.1109/ICASSP.2003.1199099.

Goubet, E., J. Katz, and F. Porikli. 2006. "Pedestrian Tracking Using Thermal Infrared Imaging." In *Defense and Security Symposium. International Society for Optics and Photonics*, 6206:62062C–62062C–12. doi:10.1117/12.673132.

Goulermas, J.Y., A.H. Findlow, C.J. Nester, D. Howard, and P. Bowker. 2005. "Automated Design of Robust Discriminant Analysis Classifier for Foot Pressure Lesions Using Kinematic Data." *IEEE Transactions on Bio-Medical Engineering* 52 (9): 1549–62. doi:10.1109/TBME.2005.851519.

Gregio, F.N. 2006. "Configuração Do Trato Vocal Supraglótico Na Produção Das Vogais Do Português Brasileiro: Dados de Imagens de Ressonância Magnética". Master, PUC/SP.

Gross, R., and J. Shi. 2001. *The CMU Motion of Body (MoBo) Database*. Robotics Institute, Carnegie Mellon University.

Guo, Y., G. Xu, and S. Tsuji. 1994. "Understanding Human Motion Patterns." In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, 1994. Vol. 2 - Conference B: Computer Vision Amp; Image Processing*, 2:325–329 vol.2. doi:10.1109/ICPR.1994.576929.

Guzman, M., A. Laukkanen, P. Krupa, J. Horáček, J.G. Švec, and A. Geneid. 2013. "Vocal Tract and Glottal Function During and After Vocal Exercising With Resonance Tube and Straw." *Journal of Voice* 27 (4): 523.e19–523.e34. doi:10.1016/j.jvoice.2013.02.007.

Hamarneh, G. 1999. "Active Shape Model Software." http://www.cs.sfu.ca /~hamarned/software/code/asm.zip.

Harshman, R., P. Ladefoged, and L. Goldstein. 1977. "Factor Analysis of Tongue Shapes." *The Journal of the Acoustical Society of America* 62 (3): 693–707. doi:10.1121/1.381581.

Höwing, F., L.S. Dooley, and D. Wermser. 1999. "Tracking of Non-Rigid Articulatory Organs in X-Ray Image Sequences." *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society* 23 (2): 59–67.

Hu, W., X. Zhou, W. Li, W. Luo, X. Zhang, and S. Maybank. 2013. "Active Contour-Based Visual Tracking by Integrating Colors, Shapes, and Motions." *IEEE Transactions on Image Processing* 22 (5): 1778–92. doi:10.1109/TIP.2012.2236340.

Huttenlocher, D.P., and W.J. Rucklidge. 1993. "A Multi-Resolution Technique for Comparing Images Using the Hausdorff Distance." In *, 1993 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1993. Proceedings CVPR '93*, 705–6. doi:10.1109/CVPR.1993.341019.

Inohara, K., Y.I. Sumita, N. Ohbayashi, S. Ino, T. Kurabayashi, T. Ifukube, and H. Taniguchi. 2010. "Standardization of Thresholding for Binary Conversion of Vocal Tract Modeling in Computed Tomography." *Journal of Voice* 24 (4): 503–9. doi:10.1016/j.jvoice.2008.10.013.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, U.K. ; New York, NY: Cambridge University Press.

Jaimes, A., and N. Sebe. 2007. "Multimodal Human–computer Interaction: A Survey." *Computer Vision and Image Understanding*, Special Issue on Vision for Human-Computer Interaction, 108 (1–2): 116–34. doi:10.1016/j.cviu.2006.10.019.

Jang, D., and H. Choi. 2000. "Active Models for Tracking Moving Objects." *Pattern Recognition* 33 (7): 1135–46. doi:10.1016/S0031-3203(99)00100-4.

Junta Doi, and M. Yamanaka. 2004. "Biometric Authentication Using Finger and Palmar Creases." In , 72–76. IEEE. doi:10.1109/VECIMS.2004.1397190.

KaewTraKulPong, P., and R. Bowden. 2002. "An Improved Adaptive Background Mixture Model for Real-Time Tracking with Shadow Detection." In *Video-Based Surveillance Systems*, 135–44. Springer US.

Kagawa, Y., Y. Ohtani, and R. Shimoyama. 1997. "Vocal Tract Shape Identification from Formant Frequency Spectra — A Simulation Using Three-Dimensional Boundary Element Models." *Journal of Sound and Vibration* 203 (4): 581–96. doi:10.1006/jsvi.1996.0865.

Karaulova, I. A, P. M Hall, and A. D Marshall. 2002. "Tracking People in Three Dimensions Using a Hierarchical Model of Dynamics." *Image and Vision Computing* 20 (9–10): 691–700. doi:10.1016/S0262-8856(02)00059-8.

Kass, Michael, Andrew Witkin, and Demetri Terzopoulos. 1988. "Snakes: Active Contour Models." *International Journal of Computer Vision* 1 (4): 321–31. doi:10.1007/BF00133570.

Kehl, Roland, and Luc Van Gool. 2006. "Markerless Tracking of Complex Human Motions from Multiple Views." *Computer Vision and Image Understanding*, Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour, 104 (2–3): 190–209. doi:10.1016/j.cviu.2006.07.010.

Kim, D., V. Maik, D. Lee, J. Shin, and J. Paik. 2006. "Active Shape Model-Based Object Tracking in Panoramic Video." In *Computational Science – ICCS 2006*, 922–29. Lecture Notes in Computer Science 3994. Springer Berlin Heidelberg.

Kim, J., A. Lammert, P. Kumar Ghosh, and S. Narayanan. 2014. "Co-Registration of Speech Production Datasets from Electromagnetic Articulography and Real-Time Magnetic Resonance Imaging." *The Journal of the Acoustical Society of America* 135 (2): EL115–EL121. doi:10.1121/1.4862880.

Klein, C. M., and J. A. Ventura. 1994. "Algorithms for Automated Inspection." *Math. Comput. Model.* 19 (11): 83–93. doi:10.1016/0895-7177(94)90018-3.

Kolahi, A., M. Hoviattalab, T. Rezaeian, M. Alizadeh, M. Bostan, and H. Mokhtarzadeh. 2007. "Design of a Marker-Based Human Motion Tracking System." *Biomedical Signal Processing and Control* 2 (1): 59–67. doi:10.1016/j.bspc.2007.02.001.

Korč, Filip, and Václav Hlaváč. 2008. "Detection and Tracking of Humans in Single View Sequences Using 2D Articulated Model." In *Human Motion*, edited by Bodo Rosenhahn, Reinhard Klette, and Dimitris Metaxas, 105–30. Computational Imaging and Vision 36. Springer Netherlands. http://link.springer.com/chapter/10.1007/978-1-4020-6693-1_5.

Koschan, A., S. Kang, J. Paik, B. Abidi, and M. Abidi. 2003. "Color Active Shape Models for Tracking Non-Rigid Objects." *Pattern Recognition Letters* 24 (11): 1751–65. doi:10.1016/S0167-8655(02)00330-6.

Krosshaug, T., J. R. Slauterbeck, L. Engebretsen, and R. Bahr. 2007. "Biomechanical Analysis of Anterior Cruciate Ligament Injury Mechanisms: Three-Dimensional Motion Reconstruction from Video Sequences." *Scandinavian Journal of Medicine & Science in Sports* 17 (5): 508–19. doi:10.1111/j.1600-0838.2006.00558.x.

Kwon, K.S., S.H. Park, E.Y. Kim, and H.J. Kim. 2007. "Human Shape Tracking for Gait Recognition Using Active Contours with Mean Shift." In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, 690–99. Lecture Notes in Computer Science 4552. Springer Berlin Heidelberg.

Laukkanen, A.M., J. Horáček, P. Krupa, and J.G. Švec. 2012. "The Effect of Phonation into a Straw on the Vocal Tract Adjustments and Formant Frequencies. A Preliminary MRI Study on a Single Subject Completed with Acoustic Results." *Biomedical Signal Processing and Control* 7 (1): 50–57. doi:10.1016/j.bspc.2011.02.004.

Li, L., W. Huang, I.Y.H. Gu, and Q. Tian. 2003. "Foreground Object Detection from Videos Containing Complex Background." In *Proceedings of the Eleventh ACM International Conference on Multimedia*, 2–10. MULTIMEDIA '03. New York, NY, USA: ACM. doi:10.1145/957013.957017.

Little, J., and J. Boyd. 1998. "Recognizing People by Their Gait: The Shape of Motion." *Videre: Journal of Computer Vision Research* 1 (2): 1–32.

Liu, G., and X. Tang. 2010. "Human Motion Tracking Based on Unscented Kalman Filter in Sports Domain." In *Kalman Filter*. InTech.

Liu, L., and G. Fan. 2005. "Combined Key-Frame Extraction and Object-Based Video Segmentation." *IEEE Transactions on Circuits and Systems for Video Technology* 15 (7): 869–84. doi:10.1109/TCSVT.2005.848347.

Ma, Zhen, Renato Natal Jorge, T. Mascarenhas, and João Manuel R. S. Tavares. 2013. "A Level Set Based Algorithm to Reconstruct the Urinary Bladder from Multiple Views." *Medical Engineering & Physics* 35 (12): 1819–24. doi:10.1016/j.medengphy.2013.05.002.

Maeda, S. 1988. "Improved Articulatory Models." *The Journal of the Acoustical Society of America* 84 (S1): S146–S146. doi:10.1121/1.2025845.

Martins, P., I. Carbone, A. Pinto, A. Silva, and A. Teixeira. 2008. "European Portuguese MRI Based Speech Production Studies." *Speech Communication*, Iberian Languages, 50 (11–12): 925–52. doi:10.1016/j.specom.2008.05.019.

Masaki, S., R. Akahane-Yamada, M.K. Tiede, Y. Shimada, and I. Fujimoto. 1996. "An MRI-Based Analysis of the English /r/ and /l/ Articulations." In *Proceedings of the Fourth International Conference on Spoken Language, 1996. ICSLP 96.*, 3:1581–1584 vol.3. doi:10.1109/ICSLP.1996.607922.

Masaki, S., Y. Nota, S. Takano, H. Takemoto, T. Kitamura, and K. Honda. 2008. "Integrated Magnetic Resonance Imaging Methods for Speech Science and Technology." *The Journal of the Acoustical Society of America* 123 (5): 3734–3734. doi:10.1121/1.2935240.

Matan, O., H.S. Baird, J. Bromley, Christopher J.C. Burges, J.S. Denker, L.D. Jackel, Y. Le Cun, et al. 1992. "Reading Handwritten Digits: A ZIP Code Recognition System." *Computer* 25 (7): 59–63. doi:10.1109/2.144441.

McInemey, T., and D. Terzopoulos. 1999. "Topology Adaptive Deformable Surfaces for Medical Image Volume Segmentation." *IEEE Transactions on Medical Imaging* 18 (10): 840–50. doi:10.1109/42.811261.

McInerney, T., and D. Terzopoulos. 2000. "T-Snakes: Topology Adaptive Snakes." *Medical Image Analysis* 4 (2): 73–91. doi:10.1016/S1361-8415(00)00008-6.

Meeds, E.W., D.A Ross, R.S. Zemel, and S.T. Roweis. 2008. "Learning Stick-Figure Models Using Nonparametric Bayesian Priors over Trees." In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, 1–8. doi:10.1109/CVPR.2008.4587559.

Metaxas, D., and S. Zhang. 2013. "A Review of Motion Analysis Methods for Human Nonverbal Communication Computing." *Image and Vision Computing*, Machine learning in motion analysis: New advances, 31 (6–7): 421–33. doi:10.1016/j.imavis.2013.03.005.

Mikić, I., M. Trivedi, E. Hunter, and P. Cosman. 2003. "Human Body Model Acquisition and Tracking Using Voxel Data." *International Journal of Computer Vision* 53 (3): 199–223. doi:10.1023/A:1023012723347.

Miller, N.A., J.S. Gregory, R.M. Aspden, P.J. Stollery, and F.J. Gilbert. 2014. "Using Active Shape Modeling Based on MRI to Study Morphologic and Pitch-Related Functional Changes Affecting Vocal Structures and the Airway." *Journal of Voice* 28 (5): 554–64. doi:10.1016/j.jvoice.2013.12.002.

Min, Junghye, and R. Kasturi. 2004. "Extraction and Temporal Segmentation of Multiple Motion Trajectories in Human Motion." In *Conference on Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04*, 118–118. doi:10.1109/CVPR.2004.64.

Moeslund, T.B., and E. Granum. 2001. "A Survey of Computer Vision-Based Human Motion Capture." *Computer Vision and Image Understanding* 81 (3): 231–68. doi:10.1006/cviu.2000.0897.

Moeslund, T.B., A. Hilton, and V. Krüger. 2006. "A Survey of Advances in Vision-Based Human Motion Capture and Analysis." *Computer Vision and Image Understanding*, Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour, 104 (2–3): 90–126. doi:10.1016/j.cviu.2006.08.002.

Mollaei, M.R.K., and M. Hassani. 2008. "Modeling of Vocal Tracts Based on Polynomials." *Computers & Electrical Engineering* 34 (6): 547–56. doi:10.1016/j.compeleceng.2007.01.004.

Muller, M., T. Roder, M. Clausen, B. Eberhardt, B. Kruger, and A. Weber. 2007. *Documentation: Mocap Database HDM05*. Tecnhical Report CG-2007-2. Universtat Bonn: Universtat Bonn.

Munhall, K. G., E. Vatikiotis Bateson, and Y. Tohkura. 1995. "X-ray Film Database for Speech Research." *The Journal of the Acoustical Society of America* 98 (2): 1222–24. doi:10.1121/1.413621.

Nadipally, M., A. Govardhan, and C. Satyanarayana. 2013. "Partial Fingerprint Matching Using Projection Based Weak Descriptor." In , 336–41. IEEE. doi:10.1109/ICSIPR.2013.6497996.

Narayanan, S., A.A. Alwan, and K. Haker. 1995. "An Articulatory Study of Fricative Consonants Using Magnetic Resonance Imaging." *The Journal of the Acoustical Society of America* 98 (3): 1325–47. doi:10.1121/1.413469.

Nascimento, J. C., M. A. T. Figueiredo, and J. S. Marques. 2005. "Segmentation and Classification of Human Activities." In *International Workshop on Human Activity Recognition and Modelling*. Oxford, Uk.

Nascimento, J. C., M.A.T. Figueiredo, and J.S. Marques. 2008. "Independent Increment Processes for Human Motion Recognition." *Computer Vision and Image Understanding* 109 (2): 126–38. doi:10.1016/j.cviu.2007.02.002.

Nguyen, T.H.D., T.C.T. Qui, K. Xu, A.D. Cheok, S.L. Teo, Z.Y. Zhou, A. Mallawaarachchi, et al. 2005. "Real-Time 3D Human Capture System for Mixed-Reality Art and Entertainment." *IEEE Transactions on Visualization and Computer Graphics* 11 (6): 706–21. doi:10.1109/TVCG.2005.105.

Nikolaidis, N., M. Krinidis, E. Loutas, G. Stamou, and I. Pitas. 2009. "Chapter 7 - Motion Tracking in Video." In *The Essential Guide to Video Processing (Second Edition)*, 175–230. Boston: Academic Press.

Ning, H., T. Tan, L. Wang, and W. Hu. 2004. "People Tracking Based on Motion Model and Motion Constraints with Automatic Initialization." *Pattern Recognition* 37 (7): 1423–40. doi:10.1016/j.patcog.2004.01.011.

Nissi Paul, S., and Y. Singh. 2014. "Survey on Video Analysis of Human Walking Motion." *International Journal of Signal Processing, Image Processing and Pattern Recognition* 7 (3): 99–122.

Norouznezhad, E., A. Bigdeli, A. Postula, and B.C. Lovell. 2008. "A High Resolution Smart Camera with GigE Vision Extension for Surveillance Applications." In *Second ACM/IEEE International Conference on Distributed Smart Cameras, 2008. ICDSC 2008*, 1–8. doi:10.1109/ICDSC.2008.4635711.

Oliveira, F.P.M., and J.M.R.S. Tavares. 2008. "Algorithm of Dynamic Programming for Optimization of the Global Matching between Two Contours Defined by Ordered Points." *Computer Modeling in Engineering*

& Sciences, Tech Science Press, 31 (1): 1–11. doi:10.3970/cmes.2008.031.001.

Oliver, N.M., B. Rosario, and A.P. Pentland. 2000. "A Bayesian Computer Vision System for Modeling Human Interactions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (8): 831–43. doi:10.1109/34.868684.

Osher, S., and J.A. Sethian. 1988. "Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations." *J. Comput. Phys.* 79 (1): 12–49. doi:10.1016/0021-9991(88)90002-2.

Paragios, N., and R. Deriche. 2000. "Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (3): 266–80. doi:10.1109/34.841758.

Parthasarathy, V., J.L. Prince, M. Stone, E.Z. Murano, and M. Nessaiver. 2007. "Measuring Tongue Motion from Tagged Cine-MRI Using Harmonic Phase (HARP) Processing." *The Journal of the Acoustical Society of America* 121 (1): 491–504.

Pentland, A., and S. Sclaroff. 1991. "Closed-Form Solutions for Physically Based Shape Modeling and Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13 (7): 715–29. doi:10.1109/34.85660.

Perkell, J. S., M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson. 1992. "Electromagnetic Midsagittal Articulometer Systems for Transducing Speech Articulatory Movements." *The Journal of the Acoustical Society of America* 92 (6): 3078–96.

Perrier, P., L.J. Boe, and R. Sock. 1992. "Vocal Tract Area Function Estimation From Midsagittal Dimensions With CT Scans and a Vocal Tract Cast: Modeling the Transition With Two Sets of Coefficients." *Journal of Speech Language and Hearing Research* 35 (1): 53. doi:10.1044/jshr.3501.53.

Piccardi, M. 2004. "Background Subtraction Techniques: A Review." In *2004 IEEE International Conference on Systems, Man and Cybernetics*, 4:3099–3104 vol.4. doi:10.1109/ICSMC.2004.1400815.

Pons-Moll, G., A Baak, T. Helten, M. Müller, H.-P. Seidel, and B. Rosenhahn. 2010. "Multisensor-Fusion for 3D Full-Body Human Motion Capture." In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 663–70. doi:10.1109/CVPR.2010.5540153.

Pontes, L., V.P. Vieira, A.A.L. Pontes, D. Curcio, and N.G. Biase. 2009. "Função de Transferência Das Vogais Orais Do Português Brasileiro: Análise Acústica Comparativa." *Brazilian Journal of Otorhinolaryngology* 75 (5): 680–84.

Poppe, R. 2007. "Vision-Based Human Motion Analysis: An Overview." *Computer Vision and Image Understanding*, Special Issue on Vision for Human-Computer Interaction, 108 (1–2): 4–18. doi:10.1016/j.cviu.2006.10.016.

Raeesy, Z., S. Rueda, J.K. Udupa, and J. Coleman. 2013. "Automatic Segmentation of Vocal Tract MR Images." In *2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, 1328–31. doi:10.1109/ISBI.2013.6556777.

Ramanan, D., D.A. Forsyth, and A. Zisserman. 2007. "Tracking People by Learning Their Appearance." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (1): 65–81. doi:10.1109/TPAMI.2007.250600.

Rani, M.P., and G. Arumugam. 2010. "An Efficient Gait Recognition System for Human Identification Using Modified ICA." *International Journal of Computer Science & Information Technology* 2 (1): 55–67.

Rathi, Y., N. Vaswani, A Tannenbaum, and A Yezzi. 2005. "Particle Filtering for Geometric Active Contours with Application to Tracking Moving and Deforming Objects." In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, 2:2–9 vol. 2. doi:10.1109/CVPR.2005.271.

Remondino, F., and A. Roditakis. 2004. "Human Motion Reconstruction and Animation from Video Sequences." In *17th International Conference on Computer Animation and Social Agents*, 347–54. Geneva, Switzerland: Computer Graphics Society.

Rius, I., J. Gonzàlez, M. Mozerov, and F.X. Roca. 2008. "Automatic Learning of 3D Pose Variability in Walking Performances for Gait Analysis." *International Journal for Computational Vision and Biomechanics*.

Rogez, G., C. Orrite, J. Martínez, and J.E. Herrero. 2006. "Probabilistic Spatio-Temporal 2D-Model for Pedestrian Motion Analysis in Monocular Sequences." In *Articulated Motion and Deformable Objects*, 175–84. Lecture Notes in Computer Science 4069. Springer Berlin Heidelberg.

Rote, G. 1991. "Computing the Minimum Hausdorff Distance between Two Point Sets on a Line under Translation." *Information Processing Letters* 38 (3): 123–27. doi:10.1016/0020-0190(91)90233-8.

Rueda, S., and J.K. Udupa. 2011. "Global-to-Local, Shape-Based, Real and Virtual Landmarks for Shape Modeling by Recursive Boundary Subdivision." In *Proceedings SPIE, Medical Imaging 2011: Image Processing*, 7962:796247–796247–13. doi:10.1117/12.878350.

Rui, Y., and P. Anandan. 2000. "Segmenting Visual Actions Based on Spatio-Temporal Motion Patterns." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2000.*, 1:111–118 vol.1. doi:10.1109/CVPR.2000.855807.

Sage, K., and S. Young. 1999. "Security Applications of Computer Vision." *IEEE Aerospace and Electronic Systems Magazine* 14 (4): 19–29. doi:10.1109/62.756080.

Saini, S., D. Rambli, S. Sulaiman, M.N. Zakaria, and S. Rohkmah. 2012. "Markerless Multi-View Human Motion Tracking Using Manifold Model Learning by Charting." *Procedia Engineering*, International Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012), 41: 664–70. doi:10.1016/j.proeng.2012.07.227.

Sandau, M., H. Koblauch, T. Moeslund, H. Aanæs, T. Alkjær, and E.B. Simonsen. 2014. "Markerless Motion Capture Can Provide Reliable 3D Gait Kinematics in the Sagittal and Frontal Plane." *Medical Engineering and Physics* 36 (9): 1168–75. doi:10.1016/j.medengphy.2014.07.007.

Sappa, A.D., N. Aifanti, S. Malassiotis, and M. Strintzis. 2005. "Prior Knowledge Based Motion Model Representation." *Electronic Letters on Computer Vision and Image Analysis* 5 (3): 55 – 67. doi:10.5565/rev/elcvia.106.

Sarkar, S., P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K. Bowyer. 2005. "The humanID Gait Challenge Problem: Data Sets, Performance, and Analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27: 162–77.

Schalkoff, R. 1989. *Digital Image Processing and Computer Vision*. John Wiley & Sons.

Schubert, M., T. Prokop, F. Brocke, and W. Berger. 2005. "Visual Kinesthesia and Locomotion in Parkinson's Disease." *Movement Disorders: Official Journal of the Movement Disorder Society* 20 (2): 141–50. doi:10.1002/mds.20281.

Schuldt, C., I Laptev, and B. Caputo. 2004. "Recognizing Human Actions: A Local SVM Approach." In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*, 3:32–36 Vol.3. doi:10.1109/ICPR.2004.1334462.

Serrurier, A., and P. Badin. 2005. "Towards a 3D articulatory model of velum based on MRI and CT images." *ZAS Papers in Linguistics (Speech production and perception: Experimental analyses and models)* 40: 195–211.

Shirai, K., and M. Honda. 1978. "Estimation of Articulatory Motion by a Model Matching Method." *The Journal of the Acoustical Society of America* 64 (S1): S42–S42. doi:10.1121/1.2004201.

Sigal, L., A. Balan, and M. Black. 2010. "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion." *International Journal of Computer Vision* 87 (1-2): 4–27. doi:10.1007/s11263-009-0273-6.

Silva, S., A. Teixeira, C. Oliveira, and P. Martins. 2013. "Segmentation and Analysis of Vocal Tract from MidSagittal Real-Time MRI." In *Image Analysis and Recognition*, 459–66. Lecture Notes in Computer Science 7950. Springer Berlin Heidelberg.

Šimšík, D., J. Majerník, A. Galajdová, and L. Želinský. 2005. "Study of Spondylolisthesis Using Videomotion Analysis." *Computer Methods in Biomechanics and Biomedical Engineering* 8 (sup1): 293–94. doi:10.1080/10255840512331389415.

Singh, S., S.A Velastin, and H. Ragheb. 2010. "MuHAVi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods." In *2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 48–55. doi:10.1109/AVSS.2010.63.

Soquet, A., V. Lecuit, T. Metens, and D. Demolin. 1996. "From Sagittal Cut to Area Function: An MRI Investigation." In *Proceedings of the Fourth International Conference on Spoken Language, 1996. ICSLP 96.*, 2:1205–1208 vol.2. doi:10.1109/ICSLP.1996.607824.

Stauffer, C., and W.E.L. Grimson. 1999. "Adaptive Background Mixture Models for Real-Time Tracking." In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999.*, 2:-252 Vol. 2. doi:10.1109/CVPR.1999.784637.

Stone, M. 1991. "Imaging the Tongue and Vocal Tract." *The British Journal of Disorders of Communication* 26 (1): 11–23.

Stone, M., Y. Cheng, and A. Lundberg. 1997. "Using Principal Component Analysis of Tongue Surface Shapes to Distinguish among Vowels and Speakers." *The Journal of the Acoustical Society of America* 101 (5): 3176–77. doi:10.1121/1.419155.

Stone, M., E. P. Davis, A. S. Douglas, M. N. Aiver, R. Gullapalli, W. S. Levine, and A. J. Lundberg. 2001. "Modeling Tongue Surface Contours from Cine-MRI Images." *Journal of Speech, Language, and Hearing Research: JSLHR* 44 (5): 1026–40.

Story, B.H. 2008. "Comparison of Magnetic Resonance Imaging-Based Vocal Tract Area Functions Obtained from the Same Speaker in 1994 and 2002." *The Journal of the Acoustical Society of America* 123 (1): 327–35. doi:10.1121/1.2805683.

Story, B.H., I.R. Titze, and E.A. Hoffman. 1996. "Vocal Tract Area Functions from Magnetic Resonance Imaging." *The Journal of the Acoustical Society of America* 100 (1): 537–54. doi:10.1121/1.415960.

Straka, M., S. Hauswiesner, M. Rüther, and H. Bischof. 2011. "Skeletal Graph Based Human Pose Estimation in Real-Time." In *Proceedings of the British Machine Vision Conference*, 69.1–69.12. BMVA Press. doi:10.5244/C.25.69.

Sundaresan, A, and R. Chellappa. 2008. "Model Driven Segmentation of Articulating Humans in Laplacian Eigenspace." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (10): 1771–85. doi:10.1109/TPAMI.2007.70823.

Szeliski, R. 2010. *Computer Vision: Algorithms and Applications*. Springer Science & Business Media.

Takemoto, H., T. Kitamura, H. Nishimoto, and K. Honda. 2004. "A Method of Tooth Superimposition on MRI Data for Accurate Measurement of Vocal Tract Shape and Dimensions." *Acoustical Science and Technology* 25 (6): 468–74. doi:10.1250/ast.25.468.

Tanaka, H., A. Nakazawa, and H. Takemura. 2007. "Human Pose Estimation from Volume Data and Topological Graph Database." In *Proceedings of the 8th Asian Conference on Computer Vision - Volume Part I*, 618–27. ACCV'07. Berlin, Heidelberg: Springer-Verlag.

Tavares, J.M.R.S., F.J.S. Carvalho, F.P.M. Oliveira, M.J.M. Vasconcelos, I.M.S. Reis, P.C.T. Gonçalves, R.R. Pinho, and Z. Ma. 2009. "Computer

Analysis of Objects' Movement in Image Sequences: Methods and Applications." *International Journal for Computational Vision and Biomechanis*, Serials Publications, 2 (2): 209–20.

Teixeira, A., R. Martinez, L.N. Silva, L.M.T. Jesus, J.C. Príncipe, and F. Vaz. 2005. "Simulation of Human Speech Production Applied to the Study and Synthesis of European Portuguese." *EURASIP Journal on Advances in Signal Processing* 2005 (9): 1435–48. doi:10.1155/ASP.2005.1435.

Teixeira, A., and F. Vaz. 2001. "European Portuguese Nasal Vowels: An EMMA Study." In *INTERSPEECH*, 1483–86.

Teixeira, A., F. Vaz, and J. C. Príncipe. 1997. "A Software Tool to Study Portuguese Vowels." In *5th European Conference on Speech Communication and Technology (Eurospeech'97)*, 5:2543–46. Rhodes, Greece.

Terzopoulos, D., and K. Fleischer. 1988. "Deformable Models." *The Visual Computer* 4 (6): 306–31. doi:10.1007/BF01908877.

Thimm, G., and J. Luettin. 1999. "Extraction Of Articulators In X-Ray Image Sequences." In *Proceedings of the European Conference on Speech Communication and Technology*, 157–60. Budapest.

Tien, Y.L., T. Kanade, and J.F. Cohn. 2000. "Dual-State Parametric Eye Tracking." In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000. Proceedings*, 110–15. Grenoble, France. doi:10.1109/AFGR.2000.840620.

Umbaugh, S.E. 2010. *Digital Image Processing and Analysis: Human and Computer Vision Applications with CVIPtools*. CRC Press.

Van der Aa, N. P., X. Luo, G. J. Giezeman, R. T. Tan, and R.C. Veltkamp. 2011. "UMPM Benchmark: A Multi-Person Dataset with Synchronized Video and Motion Capture Data for Evaluation of Articulated Human Motion and Interaction." In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 1264–69. doi:10.1109/ICCVW.2011.6130396.

Vasconcelos, M.J.M., and J.M.R.S. Tavares. 2008. "Methods to Automatically Build Point Distribution Models for Objects like Hand Palms and Faces Represented in Images." *Computer Modeling in Engineering & Sciences*, Tech Science Press, 36 (3): 213–41. doi:10.3970/cmes.2008.036.213.

Vasconcelos, M.J.M., S.M.R. Ventura, D.R.S. Freitas, and J.M.R.S. Tavares. 2010. "Using Statistical Deformable Models to Reconstruct Vocal Tract Shape from Magnetic Resonance Images." *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 224 (10): 1153–63. doi:10.1243/09544119JEIM767.

Vasconcelos, M.J.M., S.M.R. Ventura, D.R.S. Freitas, and J.M.R.S. Tavares. 2011. "Towards the Automatic Study of the Vocal Tract From Magnetic Resonance Images." *Journal of Voice* 25 (6): 732–42. doi:10.1016/j.jvoice.2010.05.002.

Vasconcelos, M.J.M., S.M.R. Ventura, D.R.S. Freitas, and J.M.R.S. Tavares. 2012. "Inter-Speaker Speech Variability Assessment Using Statistical

Deformable Models from 3.0 Tesla Magnetic Resonance Images." *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of Engineering in Medicine* 226 (3): 185–96.

Ventura, S.M.R. 2012. "A utilização da Ressonância Magnética para a Caracterização Funcional da Fala". PhD Thesis, Porto: Porto University, Faculty of Engineering.

Ventura, S.M.R., D. R. Freitas, and J.M.R.S. Tavares. 2009. "Application of MRI and Biomedical Engineering in Speech Production Study." *Computer Methods in Biomechanics and Biomedical Engineering* 12 (6): 671–81. doi:10.1080/10255840902865633.

Ventura, S.M.R., D.R.S. Freitas, and J.M.R. S. Tavares. 2011. "Toward Dynamic Magnetic Resonance Imaging of the Vocal Tract During Speech Production." *Journal of Voice* 25 (4): 511–18. doi:10.1016/j.jvoice.2010.01.014.

Ventura, S.M.R., D.R.S. Freitas, and J.M.R.S. Tavares. 2008. "Three-Dimensional Modeling of Tongue during Speech Using MRI Data." In *CMBBE 2008 - 8th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering*. Porto, Portugal.

Ventura, S.M.R., D.R.S. Freitas, and J.M.R.S. Tavares. 2010. "Imaging of the Vocal Tract Based on Magnetic Resonance Techniques." In *Computer Vision, Imaging and Computer Graphics. Theory and Applications*, 146–57. Communications in Computer and Information Science 68. Springer Berlin Heidelberg.

Ventura, S.M.R., M.J.M. Vasconcelos, D.R.S. Freitas, I.M. Ramos, and J.M.R.S. Tavares. 2011. "Speech Articulation Assessment Using Dynamic Magnetic Resonance Imaging Techniques." In *VipIMAGE 2011 - III ECCOMAS Thematic Conference on Computacional Vision and Medical Image Processing*, 225–31. Olhão, Portugal: Taylor & Francis.

Ventura, S.M.R., M.J.M. Vasconcelos, D.R.S. Freitas, I.M. Ramos, and J.M.R.S. Tavares. 2012. "Speaker-Specific Articulatory Assessment and Measurements during Portuguese Speech Production Based on Magnetic Resonance Images." In *Language Acquisition*. Nova Science Publishers, Inc., pp. 117-138.

Volna, E., and M. Kotyrba. 2013. "Vision System for Licence Plate Recognition Based on Neural Networks." In *2013 13th International Conference on Hybrid Intelligent Systems (HIS)*, 140–43. doi:10.1109/HIS.2013.6920470.

Wang, J.J., and S. Singh. 2003. "Video Analysis of Human Dynamics—a Survey." *Real-Time Imaging* 9 (5): 321–46. doi:10.1016/j.rti.2003.08.001.

Wang, L., W. Hu, and T. Tan. 2003. "Recent Developments in Human Motion Analysis." *Pattern Recognition* 36 (3): 585–601. doi:10.1016/S0031-3203(02)00100-0.

Wang, L., T. Tan, H. Ning, and W. Hu. 2003. "Silhouette Analysis-Based Gait Recognition for Human Identification." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (12): 1505–18. doi:10.1109/TPAMI.2003.1251144.

Wren, C., A. Azarbayejani, T. Darrell, and A. Pentland. 1996. "Pfinder: Real-Time Tracking of the Human Body." In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996*, 51–56. doi:10.1109/AFGR.1996.557243.

Xin, A.Y., X. Li, and M. Shah. 2004. "Object Contour Tracking Using Level Sets." In *Asian Conference on Computer Vision, ACCV 2004, Jaju Islands, Korea*.

Xue, S.A., and J.G. Hao. 2006. "Normative Standards for Vocal Tract Dimensions by Race as Measured by Acoustic Pharyngometry." *Journal of Voice: Official Journal of the Voice Foundation* 20 (3): 391–400. doi:10.1016/j.jvoice.2005.05.001.

Yoo, Jang-Hee, and Mark S. Nixon. 2011. "Automated Markerless Analysis of Human Gait Motion for Recognition and Classification." *ETRI Journal* 33 (2): 259–66. doi:10.4218/etrij.11.1510.0068.

Yoo, Jang-Hee, Mark S. Nixon, and Chris J. Harris. 2002. "Extracting Gait Signatures Based on Anatomical Knowledge." In , 596–606. London, UK: The British Machine Vision Association and Society for Pattern Recognition. http://eprints.soton.ac.uk/256502/.

Yu, S., D. Tan, and T. Tan. 2006. "A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition." In *18th International Conference on Pattern Recognition, 2006. ICPR 2006*, 4:441–44. doi:10.1109/ICPR.2006.67.

Yuille, A.L., P.W. Hallinan, and D.S. Cohen. 1992. "Feature Extraction from Faces Using Deformable Templates." *International Journal of Computer Vision* 8 (2): 99–111. doi:10.1007/BF00127169.

Zhang, Y.J. 2001. "A Review of Recent Evaluation Methods for Image Segmentation." In *Signal Processing and Its Applications, Sixth International, Symposium On. 2001*, 1:148–151 vol.1. doi:10.1109/ISSPA.2001.949797.

Zhao, M., J. Zhao, S. Zhao, and Y. Wang. 2006. "A Novel Method for Moving Object Detection in Intelligent Video Surveillance Systems." In *2006 International Conference on Computational Intelligence and Security*, 2:1797–1800. doi:10.1109/ICCIAS.2006.295372.

Zhou, H., and H. Hu. 2008. "Human Motion Tracking for rehabilitation—A Survey." *Biomedical Signal Processing and Control* 3 (1): 1–18. doi:10.1016/j.bspc.2007.09.001.