Human Body Parts Segmentation via Stacked and Multi-task Learning

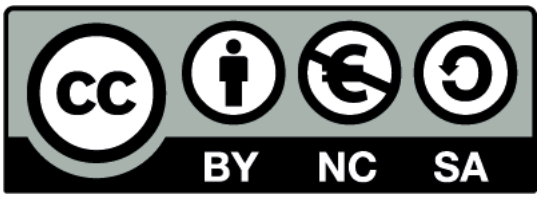
DISSERTATION

Submitted to the Doctoral Programme in Network and Information Technologies of the Universitat Oberta de Catalunya in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Author: Daniel Sánchez Abril

Thesis Director: Xavier Baró, Sergio Escalera

May 2019



Acknowledgements

My time, my space, my life. The decisions that I have been making throughout my life before starting the doctorate are as crucial as the ones I made during the research and the ones I will take when I finish. A whole series of motivations, energies, dynamics have been flowing during these years of research not in me, but in the interdependence of people who have accompanied me. From my center and with openness, I want to express my best thanks to those people who have been throughout these years.

Somewhere the spark of interest begins, that place is called University of Barcelona (UB). Specifically, the Human Pose Recovery and Behavior Analysis research group (HuPBA) led by Sergio Escalera. He is the leading companion I have had walking with me in these years of research. I wanted to learn as much as possible from him: the way to approach the problems from a skeptic point of view but at the same time opening myself to new ideas, professionalism, and respect for science, motivate me in difficult times, I still have to to learn a lot. Sergio Escalera had given me light on many occasions, at times when I wanted to escape, literally, from research and continue rowing even if everything looks dark.

My great thanks to another great mentor and researcher, Miguel Angel Bautista. Your help has been unconditional, indispensable throughout this trip. I would not be writing these lines without the great help that he gave me when I started the doctorate and motivating me in good times as well as difficult ones. He has shown me that effort is something indispensable to go far. I'm proud to have worked with you.

During part of these years, I coincided in my life obtaining a doctoral fellowship at the Open University of Catalonia (UOC). I want to thank this university for the contribution, support during this time. Thanks to the workspace in which I could go paddling, good or bad, but at least paddle.

Right at the UOC, I am happy to be part of the research group Scene Understanding and Artificial Intelligence (SUnAI). It has helped me to have met Xavier Baró during this period to guide me in the research of the doctorate. I appreciate the talks we have had and exchange ideas in these years — not only the investigation but also the personal one.

At the same time, I want to mention several people from both research groups and

the area. My warmest thanks to great people: Toni, Ágata, Albert, Miguel, Xavier, Cristina, Julio, Meysam, Marc, Ciprian, Victor, Andrei, Carles, David, Adriana, Michal, Juan Carlos, Eloi, and Oriol.

Besides, to mention the great support of the people with whom I have been able to share office at the UOC and outside, their energy has been essential so that I could be writing these lines at this precise moment. Ronak, Waseem, Hassan, Pilar, Marta, Mitch, Carles, Pedro, Amir, Fernanda, Rosen, Leila, Ania, Krizia, Eunice, Samia, Oznur, Aida, Mireia, Agnes, Marga, Juan, Tulay, Joan, Lara, Marc, Lidia, Eduardo, Raquel, Maria, Victor, Berta, Andreu, Imma, Negar, Daniel, Monika, Greig, Manuel, Maria Luisa.

Quiero tomar mis últimos agradecimientos, a mi gente más cercana. Primero no olvidar a mis grandes compañeros de vida, Aitor Marc y Julio. Caminamos juntos antes y después de todo esto, no hay palabras para describir.

Una especial mención a mi prima Gina, un referente para mí cómo persona e investigadora.

Finalmente, en mi día a día, agradecer a mis raíces, mi padre y mi madre junto a mi familia y ancestros. Ellos han sido un soporte, aguante vital junto a mi hermana Jessica. Este trabajo va dedicado exclusivamente a ellos tres.

Gente del pasado, presente y futuro.

C'est fini.

Abstract

Human Body Segmentation in RGB images has been a core problem on the Computer Vision field since its early beginnings. In this particular problem, the goal is to provide with a complete segmentation of the human/s body parts appearing in an image, discriminating the human body from the rest of the image. It is a very challenging area since it has to face many handicaps related to high variability in data such as lighting conditions, cluttering, clothes, appearance, background, point of view and number of human body parts, among others. Even so, it has become one of the areas of research because of its capabilities in real applications (i.e.surveillance, medical imaging, sign language, interactive virtual reality systems).

Hand-crafted methods covered traditional methods such as simple matching templates, deformable models, pictorial structures with tree and loopy models and discriminative ensembles learning. These approaches took researchers to point out rigorous studies to constraint the problem either by kinematic structure reasons or variability in poses/samples. However, with the appearance of deep-based methods, the traditional pipelines and methods have changed to use Deep Convolutional Neural Networks in its different variations merely. As a result, deep-based methods have been surpassing by a large margin the hand-crafted methods getting the researchers to focus on the latter methods and in their combination with traditional ones.

The writing of this thesis coincides with the paradigm shift; therefore, it is evidenced into two distinctive blocks. In the first block, we focus on a novel dataset in order to extend the state-of-the-art in human pose estimation and body segmentation. Next, we present a novel two-stage approach for human body part segmentation. We propose to use a cascade of classifiers as body parts detectors combining their outputs in an Error-Correcting Output Codes framework. Once we obtain the body pose, we apply Graph Cut segmentation optimization. Then, we use HOG features to describe the dataset and train SVM classifiers combined with the ECOC framework to feed a body part segmentation Graph Cut approach.

Moreover, we face full body segmentation, but differently, we present a novel two-stage human body segmentation method based on the discriminative Multi-Scale Stacked

Sequential Learning (MSSL) framework. In the first stage of our method for human segmentation, a multi-class Error-Correcting Output Codes classifier (ECOC) is trained to detect body parts and to produce a soft likelihood map for each body part. In the second stage, multi-scale decomposition of these maps and a neighborhood sampling is performed, resulting in a new set of features. This extensive set is trained in a stacked learning fashion with a Random Forest binary classifier. Finally, in order to obtain the resulting binary human segmentation, a post-processing step is performed through Graph Cuts optimization, which is applied to the output of the binary classifier.

In the second block of the thesis, we analyze four related human analysis tasks in still images in a multi-task scenario by leveraging synthetic datasets. Specifically, we study the correlation of 2D/3D pose estimation, body part segmentation, and full-body depth estimation. The main goal is to analyze how training together these four related tasks can benefit each task for a better generalization. Results show that all four tasks benefit from the multi-task approach, but with different combinations of tasks.

In conclusion, this thesis shows the benefit of stacked and multi-task learning for the task of human body part segmentation in still images.

Resumen

La segmentación de personas en imágenes RGB ha sido un problema central en el campo de la Visión por Computador desde sus inicios. En este problema en particular, el objetivo es proporcionar una segmentación completa de las partes del cuerpo de la persona que aparecen en dicha imagen. De esta manera, discriminando el cuerpo entero del resto de lo que aparezca en la imagen. Es una línea de investigación muy compleja, ya que se tiene que lidiar con muchos obstáculos relacionados, por ejemplo, la variabilidad de los datos, cambios de tonalidad de la iluminación, aparición de multitud de objetos en la imagen, la variedad de vestimentas por persona, apariencia física, tipo de paisaje o escenario, el sitio desde dónde se ha sacado la imagen y el número de partes del cuerpo humano, entre otros casos. Aun así, se ha convertido en una de las principales áreas de investigación debido a sus potenciales capacidades en aplicaciones (por ejemplo, video-vigilancia, tratamiento de imágenes médicas, lenguaje de signos, sistemas de realidad virtual con el que interactuar).

Los métodos diseñados manualmente han sido los métodos tradicionales tales como simples plantillas de repetición de patrones, modelos deformables, estructuras pictóricas con modelos de árboles, en bucle y aprendizaje discriminativo vía ensamblado. Todas estas variantes, llevaron a los investigadores a realizar rigurosos estudios para acotar el problema. La principal razón fue la complejidad, por ejemplo, el gran abanico de poses que la persona puede realizar. Sin embargo, con la aparición de métodos de aprendizaje profundo, los métodos tradicionales han sido substituidos por los modelos llamados redes neuronales convolucionales y sus diferentes subtipos. Como resultado, los métodos de aprendizaje profundo han superado en cierto grado a los métodos diseñados manualmente. Este hecho hace que los investigadores se centren en éstos primeros métodos y en su uso complementario con los métodos tradicionales.

En el momento de escribir esta tesis, coincide con el cambio de paradigma; teniendo esto en cuenta, se muestra en dos bloques distintos. En el primer bloque, nos centramos en un nuevo conjunto de datos para avanzar en el estado del arte en la estimación de la pose de la persona y la segmentación del cuerpo y sus partes. A continuación, presentamos un novedoso enfoque basado en dos etapas para la segmentación de partes de personas. Proponemos utilizar una cascada de clasificadores como detectores de partes del cuerpo

combinando sus salidas con un marco corrector de códigos de errores llamado ECOC. Una vez que obtenemos la postura del cuerpo, aplicamos la optimización de segmentación vía Graph Cut. Luego, usamos las características basado en el descriptor HOG para describir el conjunto de datos y entrenar un conjunto de clasificadores SVM combinados con el marco ECOC. A continuación, inicializamos un modelo gráfico para obtener la segmentación final.

Por otro lado, también tratamos la segmentación de todo el cuerpo, pero de manera diferente, presentamos un método novedoso de segmentación del cuerpo en dos etapas basado en el marco Discriminativo de Aprendizaje Secuencial Apilado a Múltiples Escalas (MSSL). En la primera etapa de nuestro método para la segmentación, un clasificador utilizado conjuntamente con un corrector de códigos de salida de corrección de errores (ECOC) de varias clases está definido para detectar partes del cuerpo y producir un mapa inicial de probabilidad para cada parte del cuerpo. En la segunda etapa, se realiza una descomposición a gran escala de estos mapas y un muestreo de regiones colindantes, lo que resulta en un nuevo conjunto de características. Este nuevo conjunto está entrenado en una forma de aprendizaje apilado con un clasificador binario. Finalmente, para obtener la segmentación binaria, se realiza una inicialización a través de la optimización de Graph Cuts, que se aplica a la salida de dicho clasificador.

En el segundo bloque de esta tesis, analizamos cuatro problemas relacionados con el análisis humano en imágenes RGB usando el paradigma de aprendizaje multi-tarea aprovechando un conjunto de múltiples datos sintéticos. En concreto, estudiamos la correlación de la estimación de la pose 2D / 3D, la segmentación de partes del cuerpo y la estimación de la profundidad de todo el cuerpo. El objetivo principal es analizar cómo la resolución conjunta de estas cuatro tareas relacionadas puede beneficiar a cada tarea para una mejor generalización. Los resultados muestran que las cuatro tareas se benefician del paradigma multi-tarea, pero combinándolas de diferentes maneras.

En conclusión, esta tesis muestra el beneficio del aprendizaje apilado y multi-tarea para el problema de segmentación de partes de la persona en imágenes.

Resum

La segmentació de persones en imatges RGB ha estat un problema central en el camp de la Visió per Computador des dels seus inicis. En aquest problema en particular, l'objectiu és proporcionar una segmentació completa de les parts del cos de la persona que apareixen en la imatge. D'aquesta manera, discriminant el cos sencer de la resta del que aparegui a la imatge. És una línia d'investigació molt complexa, ja que s'ha de tenir en compte molts obstacles relacionats, per exemple, la variabilitat de les dades, canvis de tonalitat de la il·luminació, aparició de multitud d'objectes en la imatge, la varietat de vestimentes per persona , aparença física, tipus de paisatge o escenari, el lloc des d'on s'ha tret la imatge i el número de parts del cos humà, entre altres casos. Tot i així, s'ha convertit en una de les principals àrees d'investigació a causa de les seves potencials capacitats en aplicacions (per exemple, vídeo-vigilància, tractament d'imatges mèdiques, llenguatge de signes, sistemes de realitat virtual amb el qual interactuar).

Els mètodes dissenyats manualment han estat els mètodes tradicionals com ara simples plantilles de repetició de patrons, models deformables, estructures pictòriques com models d'arbres, en bucle i aprenentatge discriminatiu via ensamblatge. Totes aquestes variants, van portar als investigadors a realitzar rigorosos estudis per delimitar el problema. La principal raó va ser la complexitat, per exemple, el gran ventall de postures que la persona pot realitzar. No obstant això, amb l'aparició de mètodes d'aprenentatge profund, els mètodes tradicionals han estat substituïts pels models anomenats xarxes neuronals convolucionals i els seus variants. Com a resultat, els mètodes d'aprenentatge profund han superat en cert grau als mètodes dissenyats manualment. Aquest fet fa que els investigadors es centrin en aquests primers mètodes i en el seu ús complementari amb els mètodes tradicionals.

En el moment d'escriure aquesta tesi, coincideix amb el canvi de paradigma; tenint en compte això, es mostra en dos diferents blocs. En el primer bloc, ens centrem en un nou conjunt de dades per avançar en l'estat de l'art en l'estimació de la postura de la persona i la segmentació del cos i les seves parts. A continuació, presentem un nou enfoc basat en dues etapes per a la segmentació de parts de persones. Proposem utilitzar una cascada de classificadors com detectors de parts del cos combinant les seves sortides

amb un marc corrector de codis d'errors anomenat ECOC. Una vegada que obtenim la postura del cos, apliquem l'optimització de segmentació via Graph Cut. Després, fem servir les característiques basat en el descriptor HOG per descriure el conjunt de dades i entrenar un conjunt de classificadors SVM combinats amb el marc ECOC. A continuació, inicialitzem un model gràfic per obtenir la segmentació final.

D'altra banda, també tractem la segmentació del cos sencer, però de manera diferent, vam presentar un mètode nou de segmentació del cos en dues etapes basat en el marc discriminatiu d'Aprenentatge Seqüencial Apilat a Múltiples Escales (MSSL). A la primera etapa del nostre mètode per a la segmentació, un classificador utilitzat conjuntament amb un corrector de codis de sortida de correcció d'errades (ECOC) de diverses classes està definit per detectar parts del cos i produir un mapa inicial de probabilitat per a cada part del cos. En la segona etapa, es realitza una descomposició a gran escala d'aquests mapes i un mostreig de regions properes, el que resulta en un nou conjunt de característiques. Aquest nou conjunt està entrenat en una forma d'aprenentatge apilat amb un classificador binari. Finalment, per obtenir la segmentació binària, es realitza una inicialització a través de l'optimització de Graph Cuts, que s'aplica a la sortida d'aquest classificador.

En el segon bloc d'aquesta tesi, analitzem quatre problemes relacionats amb l'anàlisi humà en imatges RGB usant el paradigma d'aprenentatge multi-tasca aprofitant un conjunt de múltiples dades sintètiques. En concret, estudiem la correlació de l'estimació de la postura 2D / 3D, la segmentació de parts del cos i l'estimació de la profunditat de tot el cos. L'objectiu principal és analitzar com la resolució conjunta d'aquestes quatre tasques relacionades pot beneficiar a cada tasca per a una millor generalització. Els resultats mostren que les quatre tasques es beneficien del paradigma multi-tasca, però combinant-les de maneres diferents.

En conclusió, aquesta tesi mostra el benefici de l'aprenentatge apilat i multi-tasca per al problema de segmentació de parts de la persona en imatges.

List of contributions

Contributions of the Ph.D. research

The validation of the research has been carried out with the publication of three original papers. All analysis and experimental results presented in the publications of this thesis have been produced or partially produced by the author. The list of **published papers** is the following:

1. E. Puertas, M. Bautista, D. Sanchez, S. Escalera, and O. Pujol, "Learning to segment humans by stacking their body parts," in Proc. European Conference on Computer Vision Workshop. Springer, pp. 685–697, 2014.

2. Sanchez, D., M. A. Bautista, and S. Escalera, "HuPBA 8k+: Dataset and ECOC GraphCut Based Segmentation of Human Limbs", Neurocomputing, Vol. 150, No. A, pp. 173–188, 2015.

3. Daniel Sánchez, Meysam Madadi, Marc Oliu, Xavier Baró, Sergio Escalera, Multi-task human analysis in still images: 2D/3D pose, depth map, and multi-part segmentation, Faces and Gestures, FG, 2019.

Contents

Chapter 1

Introduction

1.1 Introduction to visual human analysis

1.1.1 Human Visual System

The biological human's condition lets us observe the world around us and infer judgments. Specifically, such idiosyncrasy regard to visual perception and reasoning is the central gap between our species and the others from the Earth, making us unique through history.

Some samples can trace such a trip in our evolutionary history (Fig. 1.1). From left to right, a paleolithic cave painting of bison from Altamira cave, Spain, dated back to 20,000 years ago. This is one of our early ancestor illustrations which resembles the different type of bison used to live in the surrounding plains. At that time, we were already able to create visual art in caves with beautiful subtle paintings. That piece of art gives us some hints of the human brain functioning. First, the way we conceive in mind an object with its characteristics, color, weight, smell, shapes, sizes, appearance, location. Second, what we understand with all this information altogether in order to perceive new objects with similar characteristics. As a result, for that example, this mechanism allowed humans painting bison on that cave. This is an exercise of creating entity sense in our mind about bison and many characteristics. Painting the head, torso, hoofs, legs, tail thoroughly and so on means that humans assign semantic meaning to each bison body part. Thus, we can understand that the bison concept can be broken down from its holistic form into a composite of parts and vice versa. Similarly, the second image in Fig. 1.1 illustrates the standard Ur wooden box made during the Sumerian Civilization 4,600 years ago. It shows human figures dressing in different ways like slaves, warriors, king, farmers. Note human ability to paint in a very detailed way the different clothes and ornaments. Following the previous example, human observations on these clothes come up reasoning the concept of dressing with its characteristics and the human consisting in a set of body parts where

Figure 1.1: From left to right. Cave painting of bison from Altamira, Spain from Wikipedia (2011). War panel from Sumerian Civilization from Wikipedia (2016b). Illustration of craftworkers in ancient Egypt from Wikipedia (2016a). Artcraft of Herakles and Athena in ancient Greece from Wikipedia (2007).

each one fits with a piece of cloth. Additionally, from the last two images in Fig. 1.1 we can observe both the variability humans illustrate themselves. First, farmers working in ancient Egypt 3,500 years ago depicts more exceptional detail and different subtle poses. Second, two representative figures in the ancient Greece 2,700 years back, Herakles and Athena, depicts finer silhouettes representing in a more precise manner pose and body parts, being able to summarize actions, gestures, and behaviors from a single image.

Those few examples illustrate some light of our visual perception from the world captured in some pieces of art. This way of understanding any scene and objects around us has brought to the fore during the history. In fact, regarding our curiosity, is an appealing question done in the center of humanity.

In like manner, in nearer current times, some researchers like Blakemore, Hubel, Wiesel, back in the '60s made a similar question '¿Is the ability to see innate or acquired?' (Blakemore and Cooper, 1970; Hubel and Wiesel, 1959, 1970). In order to pull forward that question, they made different studies about how the experience can influence the brain, in concrete, the development in the visual cortex on kittens and how it is affected in their first visual environment. They demonstrated that if one of the kitten eyes is covered for a specific period during a few weeks, the cortical cells lose their input from that eye and then the other eye only influences it. This experiment manifested that kittens deprived of vision for a few months remain blind on that eye for lifelong. Consequently, they used two different cylinders, one with only vertical bars drawn inside and the second with horizontal ones. A newborn kitten was introduced in each cylinder

for a few months. Kittens that only perceived vertical lines could only see that type of pattern and not horizontal ones for lifelong and vice versa. Furthermore, they studied that neurons in the visual cortex of the kitten can be excited by moving patterns. They set up a microelectrode in some region of the kitten's visual cortex. Then, kitten's face was fixed to look at a television screen showing vertical and horizontal bars. It turned out that the electrode, recording the responses, discharged very vigorously if a vertical bar moves in front of it. Given these experiments, these neurophysiologist researchers proofed that a new visual environment is crucial for constructing visual parts of the brain since the most basic pattern such a line and more complex ones rely on environment and experience. This support the hypothesis that there are some essential aspects of visual perception acquired rather than innate. Those neurons responding to edges as orientation detectors are the bases of Gabor filter in Computer Vision (Jones and Palmer, 1987).

1.1.2 Recognizing humans from visual data

The same question can be made from that research field which is in charge to deal with how computers can understand what is taking place on digital visual information such as images and videos. It is a sub-field of Artificial Intelligence and interdisciplinary where Machine Learning and Image Analysis overlap, called Computer Vision. Some standard tasks to cope are from simple edges detection, face recognition, human body segmentation to scene understanding passing through to many more complex tasks as autonomous driving or emotion recognition. More concretely, understanding the human in visual scenes, is still an open problem (Krizhevsky et al., 2012). The human body has many degrees of freedom, just from the wrist, ankles, head, shoulder we can decompose all body parts and joints in a wide range of kinematic configurations. This leads to a coarse list of poses than implicitly we can distinguish at high accuracy, but nowadays a computer cannot do it as simple as us. Many works are facing this problem with some astonishing results (Alp Güler et al., 2018; Kocabas et al., 2018; Li et al., 2019; Tang et al., 2018; Xiao et al., 2018; Zhang et al., 2019). However, understanding and simulating the way humans recognize our body parts to infer a pose is still far to be solved with the current research knowledge. This involves various perceptual tasks such as detection, people localization, counting them, decomposed people in their semantic body categories and labeling each pixel, among others. As a result, the visual analysis of humans on standard pictures arguably implies a hard work for a machine. As an illustration, let us imagine a human appearing in a picture that is dressing a set of clothes made of different tonalities and material, in front of intense lights and bordered by the shade of forest partially occluding him. At first, glance, see Fig. 1.2, our eyes can distinguish his body parts and assign semantic meaning to each one to end up in a holistic conception

Figure 1.2: Human pose images illustrating scenarios from MPII Human Pose Dataset (Andriluka et al., 2014) showing high intraclass variability.

of himself. On the contrary, a machine must deal with the difficulties that arise from such a scenario: appearance variation due to clothing and pose which are very person-dependent, wearing different clothes in different situations, body kinematic proportions, depth information is missing, occlusions, cluttered scenes, heterogeneous scenarios and backgrounds, camera viewpoints, lighting conditions (Leung and Yang, 1995).

Given its complexity, it is critical for the visual analysis of humans to be able to detect and segment body parts. This will allow advancing in the automatic recognition and understanding of complex behaviors. Potential real applications of the automatic analysis of humans include virtual reality, video editing, intelligent vehicles, automatic product recommendation, robotics, group behavior analysis, human-computer interaction, e-commerce, action recognition, geometry ambiguities, mixed reality interfaces, animation, among others.

To understand a human body by a machine, it is needed a set of crucial stages to describe what is and what is not a human being. Thus, it comes up with the idea of paying attention to which characteristics or features make a particular human from the rest of the objects in the scene. In that sense, a feature is any piece of information that is relevant to describe and distinguish any concept from others. In concrete, for visual data, we find elementary features such as color, shape, texture, edges, points. For example, a human body consists of an own physiognomy, look, expression, shape, height, width, colors, among other features which can be beneficial to give a preliminary description. However, that rough description could not be enough to distinguish two or more humans, and there would need more precise features in order to be able not to get confused. Then, we could think that a human consists of a composite of body parts such as head, shoulders, torso, upper/lower arms, upper/lower legs, feet. Alternatively, even, we could consider joints instead of body parts since the latter by definition is a group of joints (upper leg involves the kinematic connection of hip and knee). Therefore, it is crucial to describe the human body regarding some collection of features that would discriminate more than others. Besides, these features can be chosen beforehand with some prior knowledge. For that purpose, in Computer Vision, there exist different feature descriptors that give a

connection between the pixels of a digital image and what humans understand by looking at it. Such descriptors, called handcrafted features since we specify what kind of features we want to analyze, can summarize elementary statistics information commonly used such as color, shape, regions, textures, points, edges, contours. That is, containing a low-level description of the concept to be described. For example, ones used in Computer Vision are SIFT (Lowe, 1999), HOG (Dalal and Triggs, 2005a), and SURF (Bay et al., 2006), among others.

In order to extract features, it is equally important to define which source of information is used under the feature description stage. To do so, researchers collect data regarding some predefined setting to build a dataset. This is a crucial step since the final results depend to a greater extent to the quality of the data. That quality can be interpreted from different approaches. The amount of data that is, samples are broadly collected; as a result to have a dataset as many representatives as possible. For instance, in the human body segmentation problem, it would be helpful to gather samples taking into account different sizes, heights, widths, clothing or skin color, to cite a few constraints. Then, once the features are extracted from samples, we can obtain a more representative and general description of the problem for several configurations of the human body. However, the complexity, quantity, and availability of datasets have been low till years back and probably one of the main reasons in order not to come into further advances. In concrete, human body segmentation lacked complex datasets. Existing datasets from previous years include a limited number of images, few annotations labeled on body parts, low human configuration variability and weak challenging cases like extreme poses in constrained environments. Instead, a few years back, the amount of datasets that are released, including the one introduced in this thesis, in the first block, has helped to push forward the research field. Moreover, it is worth mentioning that some of the new datasets are synthetic and work efficiently since they allow to generate automatically annotated data which one of them is evaluated in the second block of this thesis. In summary, human analysis is one of the hottest and challenging problems in computer vision. Detecting and segmenting human body and its parts in still images and image sequences (video) is an open problem, but necessary in order to define the basis for posterior human behavior understanding analysis, and to open the door to a vast range of high impact real applications.

1.1.3 Data and deep learning for human analysis

The size of available annotated datasets for human analysis has been considerably increased in recent years. This is due, in part, the large number of devices such as smartphones, sensors that generate data and massive data available on the cloud. Besides, online labeling platforms, such as Amazon Mechanical Turk also helped to provide anno-

tations to these massive amounts of available data. Under those circumstances, and taking into account the enormous computational progress in hardware during the last decade, a set of algorithms that can take profit of these conditions have started to be used by researchers. These learning algorithms mainly consist of neural networks, which take into account the difference aside, try to mimic the structured neural system embedded in the human brain.

These neural networks consist of a set of layers, each one stacked on each other, that process any data. In the case of visual data, there is a variant network called Convolutional Neural Networks (CNN) which processes images or videos. Moreover, as a general tendency, the deeper the network is in terms of layers, the highest recognition performance uses to be. This paradigm is called Deep Learning (Cireşan et al., 2011). Given the vast amount of parameters involved in deep models, i.e., tens of millions, recent progress in GPU parallel computing allowed for practical training of these models. The increase in terms of computation capability of GPUs and the amount of annotated and available data are essential in order to understand the enormous improvement in a few years by deep learning models. Equally important, Deep Learning works as a hierarchical feature extractor on the data because of its architectural nature. This makes automatically discovering and learning a particular set of features for each sample. On the contrary, handcrafted features are already selectively predefined, applied to all samples from the dataset in the same fashion and feed them to an external learning algorithm.

1.2 Objectives of the thesis

This thesis is mainly focused in human pose segmentation, but also to apply the lessons learned from human pose to multi-task learning in order to cover additional tasks such as full body depth regression and 3D pose estimation. Therefore, the main goal of this research is to find new techniques to improve human pose segmentation (both binary and multi-part) in still images, extend state-of-the-art with a new dataset and analyze the impact of dealing with multiple tasks in a multi-task learning paradigm. More precisely, we can classify the objectives of the thesis into the following goals:

1. Develop a new large and complex dataset to cover mainly human pose estimation and complementary action recognition. This dataset also serves to evaluate human body segmentation strategies.

2. Propose a two-stage segmentation approach based on the ECOC framework to evaluate human body segmentation.

3. Propose a two-stage scheme based on discriminative Multi-Scale Stacked Sequential Learning approach to tackle human body segmentation.

4. Analyze multiple human analysis tasks from a synthetic dataset in a multi-task framework.

Human segmentation in RGB images is a vital Computer Vision tasks nowadays, and it has recently attracted much interest in the research community due to its need as the first stage of many human-related applications. Nevertheless, it is an arduous task because of the full range of human poses and variability in many human patterns.

In order to provide with baseline results on the proposed dataset, we proposed a new model to benefit from ensembles of classifiers and error correction. The new ECOC-based model can contextualize several classifiers instead of merely using predictions with no context agreement.

Several approaches such as graphical models, a cascade of classifiers, generative models are studied by the community to deal with human body segmentation. A recent approach called Stacked Learning had paid attention by some researchers to benefit from the decisions of previous classifiers to be used as input features of a posterior classifier. In this case, a meta-learner is introduced in order to learn if those decisions make sense, that is, to refine them in a higher-learning level.

In a different scenario, some approaches deal with one source of information, such as 2D key-point coordinates or pixel labeling for human pose estimation or segmentation, respectively. It turns out that a different understanding can be given in order to tackle the problem. That is, instead of describing the problem as one task to solve, to define it as multiple tasks or subtasks to deal with. Thus, we could utilize a multi-task learning paradigm to explore the benefit of learning multiple end-to-end tasks, analyzing how they complement each other.

1.3 Organization of the thesis

The rest of this thesis is organized as follows:

- Chapter 2 presents briefly the kind of problem we are dealing with and a set of definitions useful to guide the reader through the thesis. In this sense, different methods used during the research are explained for a better understanding. More precisely, these definitions are sorted in two blocks: first, related to hand-crafted features and non-deep learning approaches; second, related to feature learning and indirectly to deep learning approaches such as deep neural networks.

- Chapter 3 contains two published contributions of this thesis related to non-deep learning techniques, as detailed below:
  - In Section 3.1, the first published contribution of the thesis is presented, Sánchez et al. (2015). In this chapter, a novel dataset is introduced in order to extend state of the art in human pose estimation and segmentation with minor effect on gesture recognition. Next, we present a novel two-stage approach for human body part segmentation. We propose to use a cascade of classifiers as body parts detectors combining their outputs in an Error-Correcting Output Codes framework. Once we obtain the body pose, we apply Graph Cut segmentation optimization. Then, we use HOG features to describe the dataset and train SVM classifiers combined with the ECOC framework. Moreover, a baseline for action recognition is introduced.
  - Section 3.2, presents the second published contribution of the thesis, Puertas et al. (2014). In this chapter, we present a novel two-stage human body segmentation method based on the discriminative Multi-Scale Stacked Sequential Learning (MSSL) framework. In the first stage of our method for human segmentation, a multi-class Error-Correcting Output Codes classifier (ECOC), is trained to detect body parts and to produce a soft likelihood map for each body part. In the second stage, multi-scale decomposition of these maps and a neighborhood sampling is performed, resulting in a new set of features. This extensive set is trained in a stacked learning fashion with a Random Forest binary classifier. Finally, in order to obtain the resulting binary human segmentation, a post-processing step is performed through Graph Cuts optimization, which is applied to the output of the binary classifier.
- Chapter 4 contains one contribution related to multi-task deep learning, as detailed below:
  - In Section 4.1, the third published contribution of the thesis is presented. In this chapter, we analyze four related human analysis tasks in still images in a multi-task scenario by leveraging synthetic datasets. Specifically, we study the correlation of 2D/3D pose estimation, body part segmentation, and full-body depth estimation. These tasks are learned via the well-known Stacked Hourglass module such that each of the task-specific streams shares information with the others. The main goal is to analyze how training together these four related tasks can benefit each task for a better generalization. Results on the newly released SURREAL dataset show that all four tasks benefit from the multi-task approach, but with different combinations of tasks.

- Lastly, Chapter 5 presents the concluding remarks of the thesis, in conjunction with some possible directions for future research.

Chapter 2

Background

2.1 Methods and definitions

2.1.1 Problem definition

Human Body Segmentation: It is a Computer Vision problem established in the category of Human Body Analysis (Gavrila, 1999; Leung and Yang, 1987). In concrete, Human Body Segmentation features from visual data such as images and videos. Its main goal is to acknowledge Human Body Shape from Background at different levels (Mori et al., 2004). At the most basic level, a human is defined by its contour as a result of defining a binary partition between the human body and its background. Next, in a higher abstract level, the human body decomposes into body parts such as head, torso, upper/lower arms, upper/lower legs and so on. Similarly, there is an approach per joint division where the human body resembles following a kinematic joint structure. That is, common joints such as neck, shoulders, elbows, ankles represent us. The main problem on this understanding is that joints cover a small area of the muscle to be segmented thoroughly. Thus, it may arise more difficulties than body part approach detection if the spatial context is not adequately taken into account. Furthermore, another abstract level consists of body parts groupings such as head, upper body, lower body or even further groups like hands and feet. Additionally, all those human body splitting abstractions could be combined to define a more accurate segmentation procedure. As a result, in this particular problem (Weinland et al., 2011), the goal of any of those approaches is to provide with a complete segmentation of the human body splitting appearing in an image in order to obtain the pixels belonging to the different parts of interest. Traditionally, the datasets that approach such problem are structured in binary masks per image where each mask represents a region category of the person/s.

2.1.2 Block I

Hand-Crafted Features: This kind of nontrainable feature extractor based paradigm has been used for the last decades along with well-established Machine Learning methods: SVM, Random Forest, Kernel methods, to cite a few. It is related to making use of the information properties present in the visual data extracted by some algorithms. For example, some predefined functions extract corners and edges in order to compose the representative feature vector for a particular concept. Some of its inspirations are basic algorithms including Harris Corner Detector (Harris and Stephens, 1988), Canny Edge Detector (Canny, 1987), Difference of Gaussians (DoG) (Marr and Hildreth, 1980), among others. Along the time, the number of features increased in order to solve more Computer Vision complex problems from such as lane detection to scene understanding, among others. In this sense, researchers take into account specific features such as occlusions and scale variations along with illumination. In particular, the design of hand-crafted features often involves finding the right trade-off between accuracy and computational efficiency. Besides, the accuracy can vary regarding the dataset samples. In contrast, some of these features are general-purpose, such as Gabor filtering and LBP features. Moreover, they are easy to implement and efficient for low-standard requirements. It is essential to remark; it is not defined along a trainable process as neural networks but just as a feature extractor stage.

Histogram of Oriented Gradients: Also known as **HOG**, it is a feature descriptor used in object detection (Dalal and Triggs, 2005a) which consists of counting the frequency of gradient orientation in different regions of an image. The idea behind such feature descriptor is that gradient orientation and magnitude represented in histograms can represent local object appearance and shape on an image. In general, HOG is expressed by four necessary computational steps: gradient computation, orientation binning, descriptor blocks and block normalization. First, a standard procedure is calculated over the image, the computation of gradient values. This is done by convolving a predefined set of kernels with the image in two directions, horizontal and vertical. Second, the image is divided into cells, which can be rectangular or radial, in order to calculate distributed histograms. These histograms are based on several channels that may vary depending on the problem and take into account the gradients values at pixel precision organized by cells. Third, as a way to make the descriptor robust to brightness, illumination and contrast changes, the gradient magnitude is locally normalized. This is achieved by grouping the cells in a larger cell, called blocks. These blocks, at the same time, are partially overlapped, so some cells of different blocks overlap to each other in order to influence

the illumination invariance in those block locations. Fourth, block normalization is computed following different options such as L2-norm, L1-norm, among others. Therefore, the non-normalized vector containing all histograms in a block is normalized as a result of obtaining all blocks from the descriptor concatenated representing the feature descriptor for a particular region of an image. Thus, HOG depends on three parameters: the number of cells per block, number of pixels per cell and the number of channels per cell histogram. Moreover, it offers a few advantages over other descriptors in terms of object shape information: gradients calculated on a dense grid, contrast normalization.

Grabcut: The primary purpose of this algorithm (Rother et al., 2004b) is to minimize the user interaction for foreground extraction. At the user level, first, the user draws a rectangle around the object in order to assign the outer space as background and inside the rectangle as an unknown combination distribution of foreground and background. Follow up with the user actions; these are the main constraints that the algorithm takes into account as a first solution to the problem. Moreover, the user can brush some areas of the image as background and foreground. Following up, a set of iterations are applied to the result in order to refine accuracy on foreground and background. In the light of the insider functioning, those regions that user brush either foreground or background will not change in the process of pixel labeling. Besides, regions outside rectangle are assigned as background and will not change. Instead, inside the box, the rectangle, the iterative process will decide which pixels label as background and foreground. As a result, the algorithm gives initial labeling based on user regions selection. Second, a probabilistic model, Gaussian Mixture Model (GMM) is learned to model the foreground and background color distribution. This model generates a new pixel distribution which disentangles better than previous initialization those pixels that are unknown, labeling them as probable foreground or probable background. These pixels assignation takes into account the initial user interaction at choosing those pixel regions either solid background or hard foreground in terms of color statistics. Furthermore, a graph, concretely, a Markov Random Field, is built representing the pixel distribution. Each node in the graph represents each pixel in the image. Edges represent the neighbor similarity. Moreover, two nodes are added that are the bases to the optimization algorithm coming next. These two additional nodes called Source and Sink are connected to the pixels. The former node to foreground pixels and the latter to background pixels. At the same time, each edge has a weight that represents the strength, in similarity terms, the connection between nodes pixels. Besides, Source and Sink connected to pixels have weights representing the probability belonging to these

two classes. As the next step, a 'mincut' (Boykov et al., 2001) algorithm is applied to segment the graph. That is separate foreground (Source) and background (Sink) using a minimization cost energy which prefers similar regions having the same label. This function is the summation of all weights edges that are cut. Once the cost energy is minimized as much as possible and the cut is done, all pixels connected to Source become foreground and the Sink ones become background. The process runs iteratively until convergence.

Stacked Learning: It is a set of multiple learning algorithms under the theoretical framework of ensemble learning. This type of ensembles defined as Stacked Generalization (Wolpert, 1992) builds on previous variants as references such as boosting, bootstrap aggregating, a bucket of models, among others. Stacked Generalization defines an ensemble of classifiers that are first trained by bootstrapping k-folds partitions of data in order to get the training predictions. As a result, it first generates a set of classifiers, meaning that there is an initial set of predictions. Consequently, those predictions are used to train another classifier, called combine-classifier or meta-classifier. Thus, the central insight behind it is to learn the degree of learnability that previous classifiers were able to learn. As an example, if one of the initial classifiers learns a category incorrectly, the combine-classifier could be able to learn such drifted behavior jointly with the others classifiers as a result of correcting the wrong behavior. Later on, and apart from classification, there was published a regression approach (Breiman, 1996). Moreover, a variant that takes into account context and long-range interactions are called Stacked Sequential Learning (Cohen, 2005). This variant deals with the following problems of sequential learning, namely: (a) how to capture and exploit sequential correlations; (b) how to represent and incorporate complex loss functions in contextual learning; (c) how to identify long-distance interactions and (d) how to make sequential learning computationally efficient.

Error Correcting Output Codes Framework: It is a meta-learning scheme that permits to expand any binary classifier to a multi-class case. In that sense, this framework offers a decomposition of a multi-class classification problem into simpler sub-problems. The representative ECOC meta-learning algorithm (Dietterich and Bakiri, 1994; Kong and Dietterich, 1995) is divided into two stages: the former regards learning, at such stage an ECOC encoding matrix is built to specify the combination of M binary classifiers that permit full multi-class classification. The latter stage is in charge of testing (decoding). A set of N training samples $\boldsymbol{X} = \{\boldsymbol{x_1}, ..., \boldsymbol{x_N}\}$, where each $\boldsymbol{x_i}$ belongs to a particular class $C_i \in \{C_1, ..., C_K\}$ and K defined as the number of classes, are classified regarding to the previous M

binary classifiers. These classifiers, also called dichotomizers, h_j, are used to build the ECOC matrix taking into account K. At each dichotomizer column, the binary class is split into $\{+1, 0, -1\}$, forming a KXM encoding ternary ECOC matrix $\boldsymbol{T}$. Then, when a new sample turns to be classified, each dichotomizers' output (columns of the matrix $\boldsymbol{T}$) is used to generate the codeword that is compared on each row from $\boldsymbol{T}$. At this stage, the decoding algorithm is responsible for finding the most similar class label for the test sample utilizing the outputs of the M binary classifiers. There are different decoding strategies such as minimization by Hamming distance (stems from the binary decisions) (Kong and Dietterich, 1995), Euclidean distance (Dietterich and Bakiri, 1994), loss-based metric (Escalera et al., 2008), among others. We recommend readers check Block I for further information about ECOC framework 3.1.3.3.

2.1.3 Block II

Feature Learning: This kind of trainable feature extractor based paradigm can learn data representations directly from raw information such as images, audio or any other kind of modality and detect which features or cluster of features are more worthwhile for particular tasks. Moreover, these features can be used for another similar task, which is called transfer learning. The main idea in this paradigm and taking as an example a simple neural network of two hidden layers is to discover multiple levels of representation through the model training process (LeCun et al., 2015). The lowest layers represent simple features/statistics, and higher layers give more essential discoveries such as semantic information, features that distinguish a concept from another. This, in turn, can come up with more significant robustness to intra-class variability. In case of visual information, such as images, for a CNN with a dozen of layers, lower layers represent edges, corners, bright spots, simple object/forms, and higher layers constitute growing sophisticated details of the image, such as shapes, patterns, semantics, more elaborated concepts such as table than a dull edge. It turns out that the lower layers already represent what Gabor filters or color blobs were representing a few years ago. Thus, these features make the network to be more discriminative than just a discrete bank of filters considered by classical approaches. Furthermore, these layers are possible to be used as a feature extractor for other related tasks.

Deep Learning: Deep Learning resembles a trainable hierarchical feature extractor that consists in a composite of stacked neural layers able to extract features from raw inputs and at the same time able to train in order to approximate an objective function for prediction purposes (LeCun et al., 2015). This research field, most

concretely, neural networks, stemmed from some findings between primary neural cortex and the way it can be built on hardware. Then, some research lines on the 70's defined basic neural blocks models but got stuck for a few decades. After all, it started back a few years ago because of the availability of more complex and fast hardware and availability of new large and annotated datasets. It turned out that these two factors were decisive to evolve the research on deep learning. As a result, deep networks stack many layers on top each other, showing high recognition performance but at the same time more complexity. It is important to remark that is has been tested that very deep networks not always reach better performance. In the case of visual data either images or videos, Convolutional Neural Networks (CNN) use raw images as input from low to high-level vision problems. There is also research behind input features instead of the raw images, that is, pixels intensities. Then, the network extracts the features from those images and trains without human intervention. In order to end up with a model with high generalization capability, it is desirable to have a dataset with thousands of samples, even millions, containing the whole visual variability of the problem at hands. This is one of the main drawbacks regarding Deep Learning. However, new datasets, the use of automatically generated synthetic data, and new research on semi/un-supervised learning are providing new findings in order to deal with the vast annotated data requirements of deep learning.

Convolutional Neural Networks: This type of network is the main one making use of visual information in the Deep Learning paradigm. It is a class of feed-forward artificial neural network which roughly mimics the animal visual cortex functioning. That is, the way Convolutional Neural Networks (CNN) preprocess data is based on the experiments that Hubel and Wiesel carried out (Hubel and Wiesel, 1959, 1970) with kittens. With attention to the whole structure of the network, the layer that captures the main difference with other neural networks is the Convolutional Layer. This layer is the core building block and works following the state-of-the-art algorithms related to the edge, corner detection. More precisely, the layer has a set of filters and bias, also called learnable kernels, which slide throw the entire input (i.e., an image) like in a sliding window fashion, and are multiplied by a particular piece of input region. Thus, between a kernel and an image region, a standard dot product is computed and assigned to the output as a result of generating a feature response map, one for each kernel. All these feature response maps from all kernels for a particular convolutional layer in a network form its output. Moreover, there are standard layers like pooling, fully connected, normalization, non-linear (ReLU, Sigmoid and so on) that form the network. One of its major strengths is that

kernels do not need to be hand-selected in contrast with classical filters approaches that were hand-chosen. Therefore, prior knowledge is independent that the network learns by itself.

Stacked Hourglass: This is an appropriate adjustment of a standard CNN (Newell et al., 2016c). In concrete, it is similar to an auto-encoder structure at which a set of convolutional layers plus others downsample the input up to a small feature response map followed by a set of symmetric layers to reach the same input size. Moreover, each symmetric layer is connected by skip connections in order to preserve more knowledge from very low resolution to very high resolution layers. This forms an Hourglass module which can be stacked with another Hourglass by combining the output of the former with the input of the latter to combine features. Then, it forms a stack of Hourglasses modules with intermediate supervision, that is a loss between modules helping to mitigate the vanishing gradient problem. Each module benefits from previous module outputs, refining and improving final network predictions.

Multi-Task Learning (MTL): It makes sense to analyze the 'autonomous driving' problem defining a set of tasks such as road, traffic signals, and pedestrian detection than just road detection which would give poorer results and minor understanding of the overall problem. Thus, the idea in multi-task learning for computer vision problems is to deal with multiple related tasks and train them jointly to enhance the recognition of isolated problems by the sharing of information. This can also serve to enhance the recognition performance of a higher level problem composed by several of these smaller tasks.

2.2 Related work

2.2.1 Hand-Crafted methods for human pose estimation and segmentation

Human body analysis in visual data is a challenging area since it has to face many handicaps related to high variability in data such as lighting conditions, cluttering, clothes, appearance, background, point of view, number of human body limbs. Even so, it has become one of the main interest areas of research because of its capabilities in final applications (i.e., surveillance, medical imaging, sign language, people recognition, interactive virtual reality systems).

A few years ago, a new application domain raised dealing with human analysis, customarily named "Looking at people," involving a set of main topics: human body(parts)

detection/segmentation and gesture recognition, both located in the sub-area of human pose estimation.

Human limb segmentation in RGB images has been a core problem on the Computer Vision field since its early beginnings. In this particular problem, the goal is to provide with a complete segmentation of the human/s body parts appearing in an image, discriminating the human body from the rest of the image. Usually, the human body segmentation is treated in a two-stage fashion. First, a human body part detection is performed, obtaining a large set of candidate body parts. Then, these detections are used as prior knowledge to be optimized by segmentation strategies in order to obtain the pixels belonging to the different limbs of interest.

Related to human pose estimation and segmentation, two main stages are commonly considered: body part detectors, and whole pose/segmentation inference. In the first stage, which is detection of body parts, usually weak classifiers are trained in order to obtain a soft prior of body parts (which are often noisy and unreliable). Most works in literature have used edge detectors, convolutions with filters, linear SVM classifiers, Adaboost or Cascading classifiers as in Viola and Jones (2001a). The work of Dalal and Triggs (2005b) detects the human body using a cascade of classifiers architecture with SVM and HOG features, which are the ones used in most state-of-the-art works. This is one of the main approaches used to initialize posterior pose estimation and segmentation approaches. Ramanan et al. (2005) used quadratic logistic regression on RGB features as the part detectors. Ramanan et al. (2007) detected body parts by using a tubular edge template as a detector, and convolved it with an image defining locally maximal responses above a threshold as detections. Then, they used a pictorial tree structure to infer the final pose of the human. Bourdev and Malik (2009) used body part detections in an AND-OR graph to obtain the pose estimation. Other works, have applied more robust part detectors such as SVM classifiers in Chakraborty et al. (2013); Gkioxari et al. (2013) or AdaBoost in Pishchulin et al. (2013a) trained on HOG features from Dalal and Triggs (2005a). Besides, Dantone et al. (2013) used Random Forest as classifiers to learn body parts. In spite that robust classifiers have been used, part detectors still involve false-positive and false-negatives problems given the similarity nature among body parts and the presence of background artifacts. Therefore, a second stage is usually required in order to provide an accurate segmentation.

In the second stage, once the human body pose is obtained, soft part detections are jointly optimized taking into account the nature of the human body. However, standard segmentation techniques (i.e.region-growing, thresholding, edge detection, among others.) are not applicable in this context due to the large variability of environmental factors (i.e., lightning, clothing, cluttering, among others.) and the changing nature of body textures. Nevertheless, one of the methods that have generated more attraction is the well-known

pictorial structure for object recognition Fischler and Elschlager (1973). It was redefined by Felzenszwalb and Huttenlocher (2000, 2005) in order to obtain an enriched final pose of the human body. Their proposed method represents an object that consists of hard parts linked by spring in order to give them some deformation. Thus, for example, a cat can be represented by a set of those parts belong to anatomical parts. Instead of an inanimate object, such a car can be interpreted by parts like wheels, linked by springs to the main structure. Those spring are interpreted as flexible cables that can shorten or enlarge. Imagine a tractor with high performance on damping. Their wheels will be more separated as long as tractor ride hills. So, they reformulate the method with complementary new techniques there were not in the '70s. Some works have applied an adaptation of pictorial structures using a set of joint limb marks to infer spatial probabilities. Andriluka et al. (2009) used body part detections in a boosting fashion to obtain the pose estimation. They proposed a method that covers non-rigid object detection and articulated pose estimation. Those cases are pedestrian detection, upper body estimation in TV footage and human full-body estimation in different scenarios. This work shows that such specialization may not be necessary, and proposes a general approach based on the pictorial structure framework. To do this, they focused on the appearance of body parts by approaching a densely sampled shape context descriptors and discriminatingly trained AdaBoost classifiers. In concrete, their approach uses robust generic part detectors that do not require a prefiltering search space for choosing those hypothesis candidates for final pose estimation. They compute dense appearance representations based on shape context descriptors and a boosting of classifiers to reduce as much as possible the false positives rate. That is, reducing the possible candidates to those with significant confidence. Additionally, once the classifiers are trained, it is applied a bootstrapping process to improve performance. Thus, combining those robust classifiers makes more accurate results. In this sense, the most known models for the optimization/inference of soft part priors are Poselets from Bourdev et al. (2010); Pishchulin et al. (2013a) and Pictorial Structures in Andriluka et al. (2009); Felzenszwalb and Huttenlocher (2000); Sapp et al. (2010a), both of which optimize the initial soft body part priors to obtain a more accurate estimation of the human pose and provide with a multi-limb detection. Later on, an extension was presented by Yang and Ramanan (2011, 2013) which proposed a discriminatively trained pictorial structure that models the body joints instead of limbs. A flexible mixture of parts is used in this work in order to alleviate those cases where the standard method fails mainly because of the rigidity of considered pictorial models. Given a set of parts, P, for each one there is a set of type T part that represents the normalized part but with foreshortening, orientation, and rotation in order to consider the variability of those parts. On the other hand, Wang and Mahadevan (2013) defined that a composition of the parts is a hierarchy for different combinations of pictorial struc-

ture. In this way, it is possible to consider different poses and connections of the parts. In contrast, there is a different approach that takes into account Stacked Learning for pose estimation Ramakrishna et al. (2014); Wolpert (1992), performing very similar to pictorial structure framework yet in a simple fashion. In order to apply a problem solver like pictorial structures, Stacked Learning Cohen (2005); Wolpert (1992) first learns a set of initial models to subsequently train a combine-model that involves previous predictions for a more fine-grained prediction. Some works in the literature tackle the problem of human body segmentation following or benefiting from a similar methodology to human body pose estimation. Vineet et al. (2011) proposed to use Conditional Random Fields (CRF) based on body part detectors to obtain a complete person/background segmentation. Belief propagation, branch and bound or Graph Cuts optimization are conventional approaches used to perform inference of the graphical models defined by either human body parts or person segmentation as in Hernández-Vela et al. (2012a,b); Rother et al. (2004a). Finally, methods like structured SVM or mixture of parts Yang and Ramanan (2011); Yu and Joachims (2009) can be used in order to take profit of the contextual relations of body parts. Following section reviews a few recent works dealing with human pose estimation and segmentation based on deep learning. Furthermore, a short review on MTL is also presented, with the main focus on the learning of multiple tasks related to the human body that can be beneficial for human pose estimation and segmentation.

2.2.2 Deep Learning and MTL for human pose estimation and segmentation

The use of deep-learning techniques has been a breakthrough in most Computer Vision applications, particularly Convolutional Neural Networks (CNN). It is the predominant methodology used by state-of-the-art approaches, including human analysis scenarios. In the case of human pose estimation, there has been an incremental shift from traditional approaches such as Random Forest, Bag of Visual Words (BOVW), SVM towards this hierarchical feature learnable extractor, CNN. For instance, Wei et al. (2016) developed a sequential prediction framework called 'Convolutional Pose Machines' that learned rich implicit spatial features to infer human pose estimation. Concretely, a repeated sequence of a basic CNN architecture is stacked in order to reuse the previous output heatmaps features with the input ones. As a result, this framework was able to learn long-range dependencies since the receptive field turned larger as the network made deeper. Then, Newell et al. (2016a) contributed to improving previous sequential prediction network by adding: residual modules, skip connections and an encoder-decoder shape. The resulting basic architecture was called 'Stacked Hourglass.' Each 'Hourglass' module consists of an encoder-decoder architecture with residual connections from encoder layers to corre-

sponding decoder ones. The residual module includes several convolutional layers plus skips connections. The skip connections from the encoder to decoder allow the model to fuse low-level features (e.g., edges, corners) with higher level features (e.g., semantics). This is the main network used by most human pose estimation works. Chu et al. (2017) analyzed the 'Stacked Hourglass' by plugging at its different decoder layers a set of Attention Modules (AM) and changing the standard Residual Modules (RM) by Hourglass Residual Modules (HRM). There were three different AM, from low-level local attention to high-level semantic attention: multi-resolution, multi-semantics and hierarchical visual scheme. Regarding RM, the new one contributed to increase the Receptive Field (converge earlier), made robust to scale changes and incorporated features from different scales. Chen et al. (2017b) published the first attempt of incorporating in an 'Hourglass' two Adversarial Module in order to highlight two issues: exploiting geometric constraints of joint inter-connectivity and incorporating priors about the structure of human bodies. The generator outputs visible and occluded heatmaps. Then, the discriminators were able to distinguish between real poses from fake ones (such as biologically implausible ones). Similarly, Chou et al. (2018) approached human pose estimation implementing an encoder-decoder-based discriminator (similar to 'Hourglass') based on Berthelot et al. (2017) in order to take into consideration the spatial relationships in the loss function and not just a binary decision of real or fake. Then, the discriminator can give some hints to improve the heatmaps.

Additionally, the human body part segmentation approaches have followed the same trend, turning a similar transition from well-known methods such as structured-SVM, CRF, MRF to deep hierarchical feature learning methods, CNN. For example, Luo et al. (2013) worked on parsing pedestrian images into semantic regions, such as legs, arms, body, head, and hair by training a Deep Decompositional Network (DDN). It was one of the first approaches using deep neural networks, which consists of three different hidden layers: occlusion estimation, data completion, and data transformation. These layers unify the traditional machine learning pipeline of data pre-processing in one-all-model since it directly maps low-level visual features to the label maps of body parts. Oliveira et al. (2016) collected images captured from a drone for people in disaster situations. Then, they took a pre-trained network on image classification and trained a refined Fully Convolutional Network (FCN) composed of multiple layers, where top layers combine upsampled outputs with layers from the bottom. The work of Luo et al. (2018) also studied how to tackle human parsing utilizing Adversarial Learning. In concrete, they made use of two discriminators to palliate adversarial side effects such as local and semantic inconsistency. Thus, one discriminator focused on low-resolution label map penalizing the semantic inconsistency (i.e., misplaced body parts). The other discriminator focused on multiples patches of the high-resolution label map dealing the local inconsistency

(i.e., blur and holes). Moreover, Kalayeh et al. (2018) faced human semantic parsing by extracting robust discriminative features from two very deep networks, Inception-V3 and ResNet-152. Then, they applied global average pooling to harness local visual features.

Given the need for large volumes of data to train deep learning models, there is a recent trend in learning MTL approaches. This paradigm shares information among different tasks for a better generalization while leveraging the amount of annotated data for each task. Such amount of data are publicly available for the automatic analysis of humans on works from Everingham et al. (2015); Liang et al. (2018); Lin et al. (2014); Varol et al. (2017). Related tasks include 2D pose estimation from Alp Güler et al. (2018); Andriluka et al. (2014); Gong et al. (2017); Lassner et al. (2017); Liang et al. (2018); Lin et al. (2014), body part segmentation from Alp Güler et al. (2018); Andriluka et al. (2014); Everingham et al. (2015); Lassner et al. (2017); Liang et al. (2018); Lin et al. (2014); Nie et al. (2017), human re-identification by Lassner et al. (2017), clothes parsing from Gong et al. (2017); Liang et al. (2018); Nie et al. (2017), motion/optical flow by Shahroudy et al. (2016); Zhang et al. (2013), depth estimation of Lassner et al. (2017); Varol et al. (2017), body shape model by Alp Güler et al. (2018); Varol et al. (2017), body parts shape segmentation of Varol et al. (2017), human 3D pose estimation from Ionescu et al. (2011, 2014); Mehta et al. (2017), or sign language recognition of Newell et al. (2016b), among others.

On the one hand, outstanding results have been achieved by using deep learning in tasks like the 2D pose in the wild by Alp Güler et al. (2018); Liang et al. (2018). On the other hand, the performance of other related tasks such as 3D pose, pixel-level segmentation, and human body depth estimation from RGB images still require further improvements in order to be accurately applied in real-world scenarios.

Recent approaches tend to benefit from unsupervised and cross-domain scenarios as Zamir et al. (2018b) in order to reuse data and deal with related tasks. One standard technique in this scope is the use of multi-task approaches from Everingham et al. (2015); Ionescu et al. (2011); Lin et al. (2014). Multi-task learning paradigm examined in depth by Baluja and Caruana (1995) has been shown to benefit human analysis tasks by leveraging the amount of data to be annotated since each image/video does not need a full annotation of all attributes: subsets of data can be annotated for different problems. Most importantly, while solving several tasks together, information is shared among them during training, providing them with complementary information for a better generalization.

Such multi-task works from He et al. (2017); Kokkinos (2017); Omran et al. (2018); Varol et al. (2018) tend to extend the number of tasks to better benefit from sharing knowledge within cross-domain tasks. One extreme example can be found in He et al. (2017), that extended the number of tasks to eight, not just analyzing humans but ob-

jects and animals. He et al. (2017) developed a pyramid image decomposition as input to deal with semantic/boundary/object detection, standard estimation saliency/normal estimation, semantic/human part segmentation, semantic boundary detection, and region proposal generation. Other works such as as Dai et al. (2016); Luvizon et al. (2018); Popa et al. (2017); Zhao et al. (2018) added additional tasks such for instance segmentation, multi-human parsing, and mask segmentation. As an example, Dai et al. (2016) faced instance segmentation, object detection and mask segmentation in a stacked fashion. Moreover, some research has been conducted by using multi-task of 2D/3D pose and body parts parsing by Alp Güler et al. (2018); Xia et al. (2017), sometimes including additional tasks as 3D body shape estimation from Omran et al. (2018); Varol et al. (2018).

One can find different strategies in order to define multi-task schemes. Zamir et al. (2018b) performed the most large-scale analysis of cross-domain for indoor scenes with no-human interaction in their new dataset. They trained 26 neural networks, one per category and new combinations related to multiple domains via transfer learning instead of multi-tasking. Most patterns found on this dataset exclude human kinematic constraints. Xia et al. (2017) built a two-stage FCN process that initially detects human pose and finally refines body parts parsing through the conditional random field. The work of Alp Güler et al. (2018) used Mask-RCNN from He et al. (2017) in a multi-task cascade fashion connecting several intermediate layers for pose estimation and body parts parsing, while Kokkinos (2017) used Mask R-CNN for instance/mask segmentation and object/key-point detection problems. The work of Zhao et al. (2018) made use of adversarial networks in a nested way, i.e., GANs outputs are used as the input to other GAN to deal with pose estimation and body parts parsing. In Popa et al. (2017); Wei et al. (2016) recursive processing stages are used to detect and segment 2d/3d pose and body-parts. While Kocabas et al. (2018) performed faster inference facing person/keypoint detection, person segmentation and pose estimation on two streams: key-point and person detection.

Another common combination of tasks is 2D/3D pose and body/clothes parsing in Ionescu et al. (2011) on datasets such as Pascal from Everingham et al. (2015) or COCO from Lin et al. (2014). The work of Nie et al. (2018) used two encoders (2D pose and clothes parsing) with a module as a middle stream that acts as a parameter adapting to merge the features of both tasks and perform classification separately. In contrast, Liang et al. (2018) proposed a two-stage multi-task procedure that first extracts sharing features with residual networks to be used in a second stage consisting of two CNN performing 2D pose estimation and clothes parsing, respectively.

Finally, Varol et al. (2017) published SURREAL, a dataset of sequences of realistic synthetic human bodies. The dataset includes RGB, 2D/3D joints, segmented body parts, optical flow, and depth information. This new dataset allows exploring new multi-task approaches for human body analysis.

Chapter 3

Block I

In this chapter, we present two distinct approaches to tackle human multi-limb and full-body segmentation. These approaches are based on traditional state-of-the-art methods such as Random Forest, AdaBoost and SVM's. Besides, we make use of hand-crafted features, in particular, HOG and Haar-like features. Complementary, methods from graphical models (GraphCuts), ensemble learning (Stacked Generalization Learning, ECOC framework) are used to optimize the learning procedure.

3.1 HuPBA 8k+: Dataset and ECOC-GraphCut based Segmentation of Human Limbs

3.1.1 Introduction

Human body analysis in visual data is a challenging area since it has to face many handicaps related to high variability in data such as lighting conditions, cluttering, clothes, appearance, background, point of view, number of human body limbs. Even so, it has become one of the main interest areas of research because of its capabilities in final applications (i.e., surveillance, medical imaging, sign language, people recognition, interactive virtual reality systems).

In the last years, a new application domain has raised dealing with human analysis, customarily named "Looking at people," involving a set of main topics that cover this work: human body(parts) detection/segmentation and gesture recognition, both located in the sub-area of human pose estimation.

Human limb segmentation in RGB images has been a the core problem in the Computer Vision field since its early beginnings. In this particular problem, the goal is to provide with a complete segmentation of the human/s body parts appearing in an image, discriminating the human body from the rest of the image. Usually, human body segmen-

tation is treated in a two-stage fashion. First, a human body part detection is performed, and then, these detections are used as prior knowledge to be optimized by segmentation strategies in order to obtain the pixels belonging to the different limbs of interest.

Related to human pose estimation, either detection or segmentation approach. In the first stage, Dalal and Triggs (2005b) detect the human body using a cascade of classifiers architecture with SVM and HOG features. This is one of the main approaches used to initialize posterior pose estimation and segmentation approaches. Agarwal and Triggs (2006) make use of human silhouettes since most of the body pose information remains there and applied regression of joint angles against clustering of a histogram of shape context. As a result, no body model or labeled localizations of body parts are needed. Ramanan et al. (2007) detect body parts and use a pictorial tree structure to infer the final pose of the human. Bourdev and Malik (2009) use body part detections in an AND-OR graph to obtain the pose estimation. Similarly, Andriluka et al. (2010) use Adaboost Classifiers as boosted part detectors and shape context representation in a tree pictorial structure to initialize a pedestrian tracking system. A distinct paradigm can be found in the work of Yao and Fei-Fei (2010) that defines a new descriptor including logical relations where local features are codified using logical operators, permitting a discriminative understanding of the person and the context. Wang et al. (2011) employ hierarchical poselets as part-based models to deal with non-rigid parts (e.g., ankle, neck, wrist) and to capture different granularity of details. Finally, Pishchulin et al. (2013a,b) expand a tree-structure conditioned on poselet hypotheses as medium level feature representation to keep an exact yet tractable inference for both unary and binary terms.

In the second stage, once the human body pose is obtained, many methods can be applied in order to obtain a human/background segmentation. Vineet et al. (2011) propose to use Conditional Random Fields based on body part detectors to obtain a complete person/background segmentation. Shotton et al. (2013) build a real-time system so-called Kinect by using depth images to train very deep random forests with depth pixel difference features for body part segmentation. Besides, there are different approaches to obtain either a multi-limb or a complete human body segmentation. One of the methods that are giving superior performance is the well-known pictorial structure in Andriluka et al. (2009); Sapp et al. (2010a) for object recognition. This method was introduced by Fischler and Elschlager (1973) and revisited by Felzenszwalb and Huttenlocher (2005), which uses a set of joint limb marks involving a pictorial structure to infer spatial probabilities. The method of Felzenszwalb and Huttenlocher (2005) represents an object that consists of hard parts linked by spring in order to give them some deformation. Andriluka et al. (2009) use body part detections in a boosting fashion to obtain the pose estimation, which proposes a method that covers non-rigid object detection and articulated pose estimation. Felzenszwalb et al. (2010) introduce mixtures of multi-scale deformable part

models where each human body part is trained discriminatively and improve matching deformable models. An extension is presented by Yang and Ramanan (2011, 2013) which proposes a discriminatively trained pictorial structure that models the body joints instead of limbs. The work of Sapp et al. (2010b) defines a discriminative coarse-to-fine cascade of pictorial structure to reduce the pose search space and get finer poses including multiple features descriptors such as contour, geometry, shape, and appearance. On the other hand, Wang and Mahadevan (2013) define that a composition of the parts is a hierarchy for different combinations of pictorial structures. In this way, it is possible to consider a different set of poses and connections of the parts. Similarly to pictorial structure, Felzenszwalb and McAllester (2011) and Girshick et al. (2011) formalize grammar models in order to provide a flexible framework for people detection and segmentation where the human body is a compositional structure of body parts complemented with deformation rules that allow relative body part movement. A compositional AND-OR graph grammar model from Rothrock et al. (2013) include the Background cue to deal with clutter scenes, occlusions in the human body and body part segmentation. Similarly, Ladicky et al. (2013) combine the articulated poses from a pictorial structure part-based model and learn a graphical structure pixel-based model plus color and texture features to segment body parts. In contrast, there are some works using Graph Cuts optimization such in human body parts segmentation of Hernández-Vela et al. (2012b) or person segmentation of Rother et al. (2004a). Eichner et al. (2012) make use of an edge-template model to learn priors of Foreground and the human body and initialize a following GrabCut procedure of Foreground/Background labeling in order to reduce body parts search space. Another kind of pose initialization is done by Sapp et al. (2011) for human motion where decoupling a complex problem in an ensemble-based approach of tree-structure per joint is learned to obtain an exact inference. Besides, there are other approaches such as the one proposed by Ramanan (2006) which uses an iterative parsing procedure for learning a model for each sample. Besides, the multi-person pose estimation from Eichner and Ferrari (2010) utilizes an upper-body detector as first rough estimates in order to incorporate a multi-pictorial structure to obtain all poses jointly with occlusion priors and an inter-people exclusion penalty.

In gesture recognition, there exists a vast number of methods based on dynamic programming algorithms for alignment and clustering of temporal series like Zhou et al. (2013). Other probabilistic methods such as Hidden Markov Models (HMM) or Conditional Random Fields (CRF) have been commonly used in the literature as in Starner and Pentland (1997). Nevertheless, one of the most common methods for Human Gesture Recognition is Dynamic Time Warping(DTW) is the one from Reyes et al. (2011) since it offers a simple yet effective temporal alignment between sequences of different lengths. Typically, in order to apply an evaluation procedure, these methods are applied on RGB

images which are collections of people performing different gestures. Therefore, part of the performance of those methods involves creating a dataset robust enough to deal with the constraints of the problem to address.

In the Computer Vision community, we can find different datasets according to variable scenarios, people, illumination characteristics and so on. Such datasets like Parse from Ramanan (2006), Buffy in Ferrari et al. (2008), UIUC People from Tran and Forsyth (2010), Pascal VOC in Everingham et al. (2010), to cite a few, are widely used to evaluate the different methods than the community use. As a result of the lack of variety of samples, we introduced by previous search, a new dataset named HuPBA in order to tackle much more specific human body analysis and recovery (i.e., multi-limb segmentation, gesture recognition). Such datasets on literature do not scale enough the number of properties such as several limbs annotation at pixel precision, limbs labeled with a second plus of gesture recognition approach.

In this chapter, we present a novel double two-stage approach for the segmentation of the human body on RGB data. We propose to use a cascade of classifiers as body part detectors in a tree-structure combining their outputs in an Error-Correcting Output Codes framework. Once the body-pose estimation is obtained, it is used as initialization of a GMM color modeling and posterior binary Graph Cut segmentation optimization. Then, HOG features are used to describe the dataset and used train SVM classifiers according to a tree-structure without taking into account the background category (since in the previous step we remove it). After that, the binary person segmentation from Graph Cut is applied to each RGB image as an overlapping base in order to constrain the region to evaluate. Once it is done, GraphCut multi-limb is applied to each image and with the priors of each limb in order to segment as many limb categories as are defined. Besides, gesture recognition is applied by using HMM and DTW. Furthermore, we provide a novel dataset consisting of 8 000 images in which 14 limbs were manually tagged. As a result of our double two-stage segmentation methodology, we show performance in comparison to state-of-art methods applied to binary segmentation, multi-limb segmentation and gesture recognition.

3.1.2 HuPBA 8K+ Dataset

Automatic human-limb detection and segmentation, human pose recovery and behavior analysis are challenging problems in computer vision, not only for the intrinsic complexity of the tasks, but also the lack of large public and annotated datasets. Usually, public available dataset lack of refined labeling or contain a very reduced number of samples per limb (e.g., *Buffy Stickmen V*3.01, *Leeds Sports* and *Hollywood Human Actions* from Ferrari et al. (2008); Johnson and Everingham (2010); Laptev et al. (2008)). Besides, large

datasets often use synthetic samples or capture human limbs with sensor technologies such as *MoCap* in very controlled environments from De la Torre et al. (2008).

Being aware of this lack of publicly available datasets for multi-limb human pose detection, segmentation and gesture recognition, we present a novel fully limb labeled dataset, the *HuPBA* $8k+$ dataset. This dataset is formed by more than 8 000 frames where 14 limbs are labeled at pixel precision[1]. Furthermore, the HuPBA $8k+$ dataset also contains gesture annotations for 11 separate and collaborative gesture categories. The main characteristics of the dataset are the following:

1. The images are obtained from 9 videos (RGB sequences) and a total of 14 different actors appear in those 9 sequences. In concrete, each sequence has the main actor (9 in total) which during the video interacts with secondary actors performing a set of different actions.

2. Each video (RGB sequence) was recorded with a 15 fps rate.

3. RGB images were stored with resolution 480x360 in BMP file format.

4. For each image 14 limbs were manually tagged: Head, Torso, R-L Upper-arm, R-L Lower-arm, R-L Hand, R-L Upper-leg, R-L Lower-leg, R-L Foot.

5. Limbs are manually labeled using binary masks, and the minimum bounding box containing each subject is defined.

6. The actors appear in a wide range of different poses and performing different actions/gestures.

7. For each video we manually labeled a set of 11 gesture categories: Wave, Point, Clap, Crouch, Jump, Walk, Run, Shake Hands, Hug, Kiss, Fight.

Finally, the easy and challenging aspects of the HuPBA $8k+$ dataset are listed in Table 3.1.

3.1.2.1 Data Format and Structure

The dataset we introduce is composed of RGB images, labeled limbs (binary masks) and additional information that has a specific structure to distinguish the location of limbs and gestures for each actor. Additionally, for each actor, a pair of structured files are created to store the location of the bounding-boxes for each RGB image and the start-end frames associated with the gestures executed. The folder structure that contains the HuPBA $8k+$ dataset is shown in Fig. 3.1.

[1]The whole number of manually labeled limbs exceeds 120 000.

Easy
Fixed Camera
Frontal point of view
Full body capture
The main actor is kept within a sequence
Several instances of each gesture
Gestures differentiated by an idle pose
Fixed background across all video sequences
Challenging
Within each sequence:
Gestures executions involve most limbs
Gestures imply the interaction of various actors
Between sequences:
Variations in clothing, skin color, height and width person
Some parts of the body may be occluded

Table 3.1: Easy and challenging aspects of the HuPBA $8k+$ dataset.

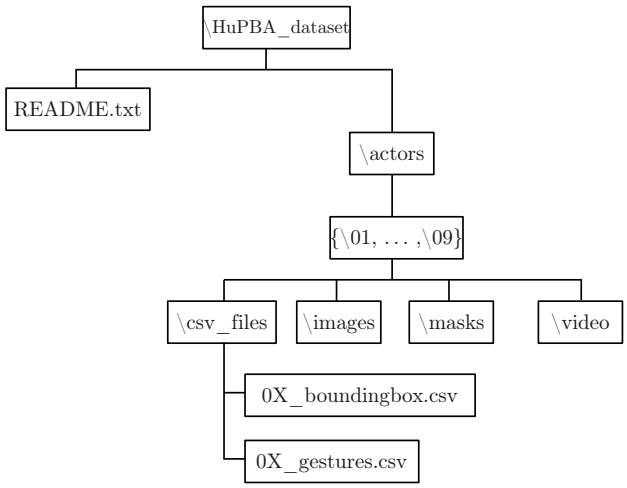


Figure 3.1: Folders structure.

3.1.2.1.1 Folder \images

In this folder, we store the set of frames for a given video sequence. The folder *\images* contains the sequence of RGB images (480x360 pixels). Each image name has the structure *idActor_numberFrame.bmp*, where:

- **idActor**: Numerical identifier of the actor $\{01, 02, ..., 09\}$.
- **numberFrame**: Numerical identifier of the image in the sequence.

3.1.2.1.2 Folder \masks

This folder contains the binary masks for each one of the 14 limbs appearing on each frame. In the case of two actors appearing in a frame, there will be a *id* for each one in order to distinguish limbs. Each binary mask name has the structure *idActor_numberFrame_idUser_idLimb.bmp*, where:

- **idActor**: Numerical identifier of the actor $\{01, 02, ..., 09\}$.
- **numberFrame**: Numerical identifier of the image in the sequence.
- **idUser**: Numerical identifier for the actor that appears in the image. Values $\{1, 2, ..., n\}$. In the case of appearing two actors: The main actor and another, the main actor is 1, the second is 2, and so on.
- **idLimb**: Numerical identifier of the limb, which are described in Fig. 3.2.

3.1.2.1.3 Bounding-boxes

In addition, for each sequence of images there is a file *0X_boundingbox.csv* located in the directory *\csv_files* that contains the bounding-boxes of all actors that appear in that sequence. That is, for each actor that appears in an image, its bounding-box is given. In the case of two actors appearing in an image, two bounding-boxes will be described, one for each actor, as shown in Fig. 3.3. The *csv* file contains the following structure:

- **id_user**: Numerical identifier for the actor that appears in the image. Values $\{1, 2, ..., n\}$. In the case of appearing two actors: The main actor and another, the main actor is 1 and the second is 2. Thus, there will be two bounding-boxes, one for 1, another for 2, and so on.
- **number_frame**: Numerical identifier of the image in the sequence.
- **x**: Minimum position of X. That is, the leftmost.

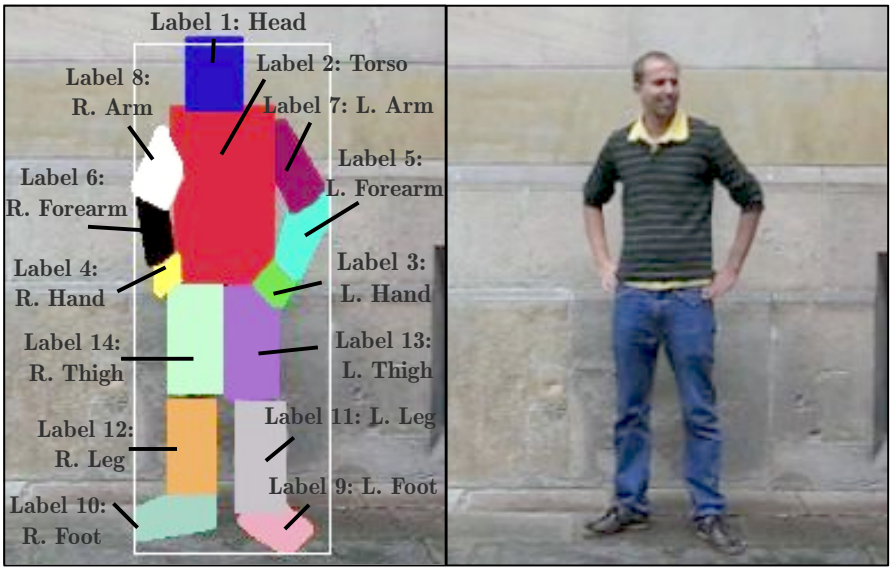


Figure 3.2: Human-Limb labelling on the HuPBA $8k+$ dataset.

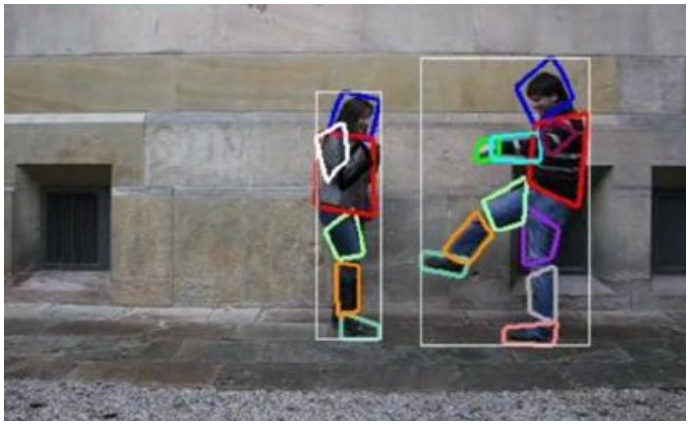


Figure 3.3: Sample of two bounding-boxes in a frame.

- **y**: Minimum position of Y. That is, the uppermost.
- **width**: Width of the bounding-box.
- **height**: Height of the bounding-box.

3.1.2.1.4 Gestures

Besides of the human-limb labeling provided on the dataset, we also annotated gestures performed by the actors. The 11 gesture categories labeled are the following: Wave, Point, Clap, Crouch, Jump, Walk, Run, Shake Hands, Hug, Kiss, and Fight. An example of keyframes for the different gesture categories are shown in Fig. 3.4. Each set of gestures

	HuPBA	PARSE 2006	BUFFY 2008	UIUC 2010	LEEDS 2010	HW 2008	MMGR13 2013	Human Actions 2004	Pascal VOC 2010
Labeling at pixel precision	Yes	No	No	No	No	-	No	No	Yes
Number of limbs	14	10	6	14	14	-	16	-	5
Number of labeled limbs	124 761	3 050	4 488	18 186	28 000	-	27 532 800	-	8 500
Number of frames	8 234	305	748	1 299	2 000	-	1 720 800	-	1 218
Full body	Yes	Yes	No	Yes	Yes	-	Yes	Yes	Yes
Limb annotation	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Gesture annotation	Yes	No	No	No	No	Yes	Yes	Yes	No
Number of gestures	11	-	-	-	-	8	20	6	-
Number of gesture samples	235	-	-	-	-	430	13 858	600	-

Table 3.2: Comparison of public dataset characteristics.

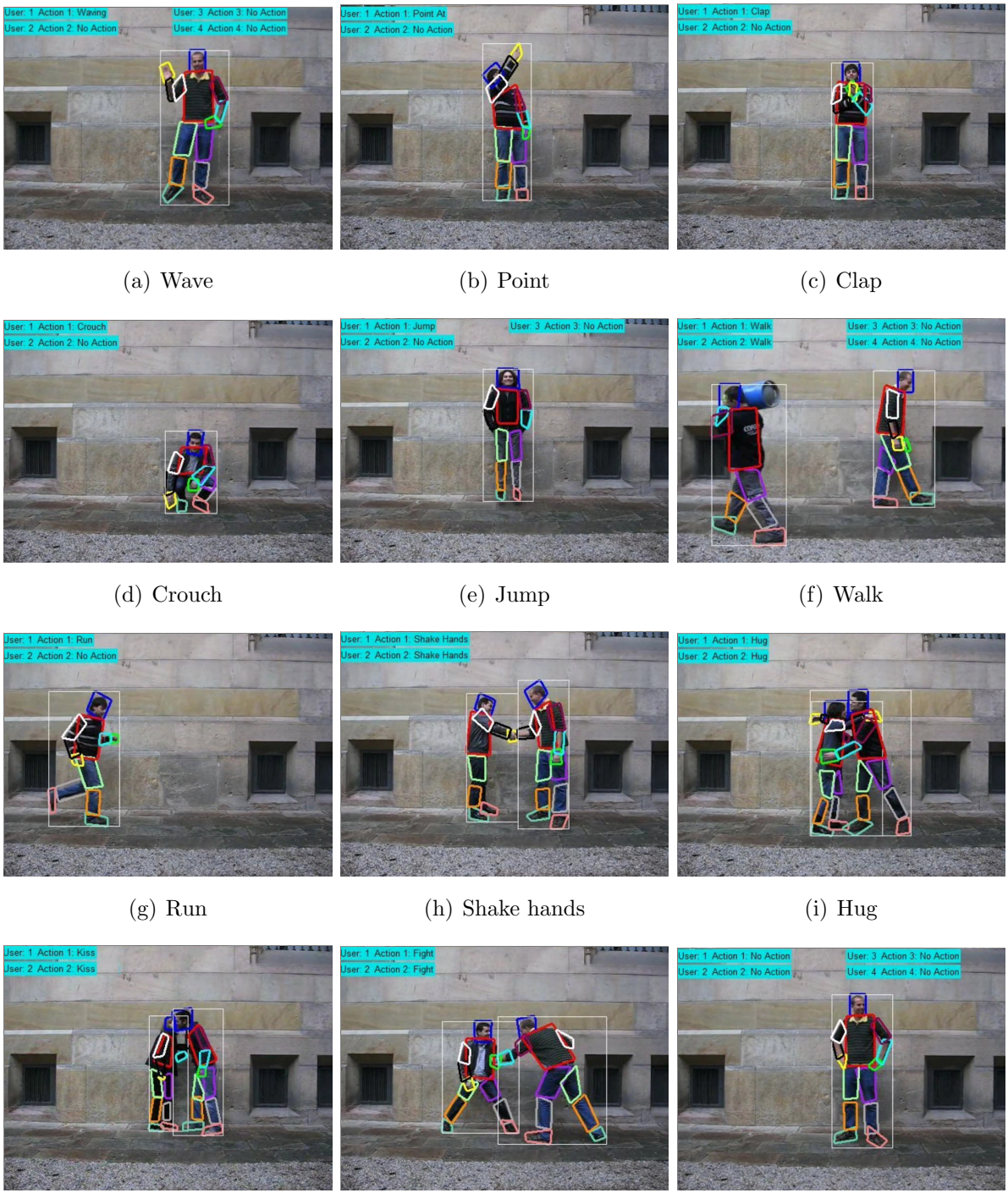
performed by an actor is associated with a file *./csv_files/0X_gestures.csv* that contains the following structure:

- **id_user**: Numerical identifier for the actor that appears in the image. Values $\{1, 2, ..., n\}$.
- **label_gesture**: Numerical identifier related to the gesture performed. There are gestures that involve just one actor (i.e. walk or run), and others more than one actor (i.e. fight or kiss).
- **start_frame**: The number of image where the gesture starts.
- **end_frame**: The number of the image where the gesture ends.

Finally, in Table 3.2 we compare the HuPBA $8k+$ dataset characteristics with some publicly available datasets. These public datasets are chosen to take into account the variability of limbs and gestures. Thus, we present a novelty dataset in which the limbs are labeled at pixel precision with more labeled limbs for many images higher than most public datasets (i.e., Pascal VOC, PARSE, BUFFY, UIUC people, LEEDS SPORTS). In case of gestures, there is more equality in the number of gestures set with the others datasets (i.e., HOLLYWOOD (HW), MMGR13, Human Actions) but ours lets work with much more precision because of limbs labeled at pixel precision. In contrast, MMGR13 present much more variety of gestures and samples than us.

3.1.3 Methodology

In the following subsections, we describe the proposed system for automatic segmentation of human limbs. To accomplish this task, we start by defining a framework divided into



(a) Wave (b) Point (c) Clap

(d) Crouch (e) Jump (f) Walk

(g) Run (h) Shake hands (i) Hug

(j) Kiss (k) Fight (l) Idle

Figure 3.4: Different gesture categories labeled on the HuPBA $8k+$ dataset. Images from (a) to (g) illustrate single actor gestures, and images from (h) to (k) show gestures that required interacting with a secondary actor. Additionally, (l) shows an example of an idle gesture.

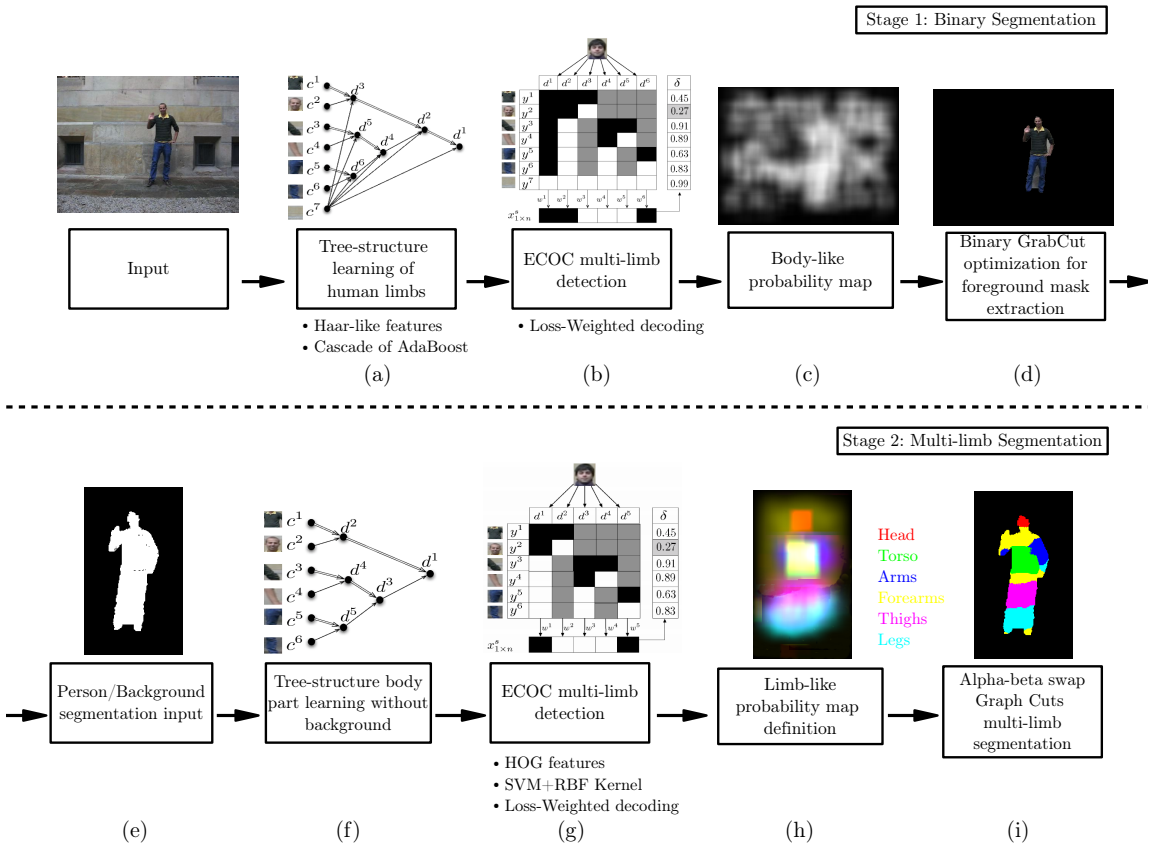


Figure 3.5: Scheme of the proposed human-limb segmentation method.

a two-stage procedure. The first stage focused on binary person/background segmentation is split in four main steps: a) Body part learning using a cascade of classifiers, b) Tree-structure learning of human limbs, c) ECOC multi-limb detection, also, d) Binary GrabCut optimization for foreground extraction. In the second stage, we segment the person/background binary mask into different limb regions. This stage is split into the following four steps: e) Tree-structure body part learning without background, f) ECOC multi-limb detection, g) Limb-like probability map definition, and h) Alpha-beta swap Graph Cuts multi-limb segmentation. The scheme of the proposed system is illustrated in Fig. 3.5.

3.1.3.1 Body part learning using a cascade of classifiers

The core of most human body segmentation methods in the literature relies on body part detectors. In this sense, most part detectors in literature follow a cascade of classifiers architecture as in Chen and Chen (2008); Enzweiler and Gavrila (2009); Freund and Schapire (1995); Mikolajczyk et al. (2004); Zhu et al. (2006). The Cascades of classifiers are based on the idea of learning and unbalanced binary problem by using the negative

outputs of a classifier d^i as an input for the following classifier d^{i+1}. Mainly, this cascade structure allows any classifier to refine the prediction by reducing the false positive rate at every stage of the cascade. In this sense, we use AdaBoost as the base classifier in our cascade architecture.

Besides, in order to make the body part detection rotation invariant, all body parts are rotated to the dominant gradient region orientation. Then, Haar-like features are used to describe body parts.

Because of its properties, a cascade of classifiers is usually trained to split one visual object from the rest of the possible objects of an image. This means that the cascade of classifiers learns to detect a particular object (body part in our case), ignoring all other objects (all other body parts). However, if we define our problem as a multi-limb detection procedure, some body parts are similar in appearance, and thus, it makes sense to group them in the same visual category. Because of this reason, we propose to learn a set of a cascade of classifiers where a subset of limbs are included in the a positive set of a cascade and the remaining limbs are included as negative instances together with background images in the negative set off the cascade. Applying this grouping for different cascades of classifiers in a tree-structure way and combining them in an Error-Correcting Output Codes (ECOC) framework enables the system to perform multi-limb detection as in Escalera et al. (2010a).

3.1.3.2 Tree-structure learning of human limbs

The first issue to take into account when defining a set of cascades of classifiers is how to define the groups of limbs to be learned by each cascade. For this task, we propose to train a tree-structure a cascade of classifiers. This tree-structure defines the set of meta-classes for each dichotomy (a cascade of classifiers) taking into account the visual appearance of body parts, which has two purposes. On the one hand, we aim to avoid dichotomies in which body parts with different visual appearance belong to the same meta-class. On the other hand, the dichotomies that deal with classes that are difficult to learn (body parts with similar visual appearance) are defined taking into account a few classes. An example of the body part tree-structure defined taking into account these issues for a set of 7 body limbs is shown in Fig. 3.6(a). Notice that classes with similar visual appearance (e.g., upper-arm and lower-arm) are grouped in the same meta-class in most dichotomies. Besides, dichotomies that deal with severe problems (e.g., d^5) are focused only on the problematic classes, without taking into account all other body parts. In this case, class c^7 denotes the background.

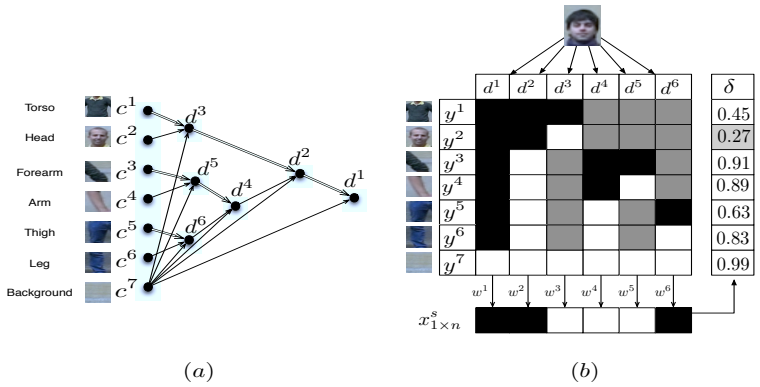


Figure 3.6: (a) Tree-structure classifier of body parts, where nodes represent the defined dichotomies. Notice that the single or double lines indicate the meta-class defined. (b) ECOC decoding step, in which a head sample is classified. The coding matrix codifies the tree-structure of (a), where black and white positions are codified as $+1$ and -1, respectively. c, d, y, w, X, and δ correspond to a class category, a dichotomy, a class codeword, a dichotomy weight, a test codeword, and a decoding function, respectively.

3.1.3.3 ECOC multi-limb detection

In the ECOC framework, given a set of N classes (body parts) to be learned, n different bi-partitions (groups of classes or dichotomies) are formed, and n binary problems over the partitions are trained as in Bautista et al. (2012b). As a result, a codeword of length n is obtained for each class, where each position (bit) of the code corresponds to a response of a given classifier d (coded by $+1$ or -1 according to their class set membership, or 0 if a particular class is not considered for a given classifier). Arranging the codewords as rows of a matrix, we define a *coding matrix* M, where $M \in \{-1, 0, +1\}^{N \times n}$. During the *decoding* (or testing) process, applying the n binary classifiers, a code x is obtained for each data sample ρ in the test set. This code is compared to the base codewords $(y^i, i \in [1, .., N])$ of each class defined in the matrix M and the data sample is assigned to the class with the *closest* codeword as in Escalera et al. (2010a).

The ECOC coding step has been widely tackled in the literature either by predefined or problem-dependent strategies. However, recent works showed that problem-dependent strategies could obtain high performance by focusing on the idiosyncrasies of the problem, similar to Bautista et al. (2014). Following this fashion, we define a problem dependent coding matrix in order to allow the inclusion of cascade of classifiers and learn the body parts. In particular, we propose to use a predefined *coding* matrix in which each dichotomy is obtained from the body part tree-structure described in the previous section. Fig. 3.6(b) shows the coding matrix codification of the tree-structure in Fig. 3.6(a).

3.1.3.3.1 Loss-weighted decoding using cascade of classifier weights

In the ECOC *decoding* step an image is processed using a windowing method, and then, each image patch, that is, a sample ρ is described and tested. In this sense, each classifier d outputs a prediction whether ρ belongs to one of the two previously learned meta-classes. Once the set of predictions $x^{\rho}_{1\times n}$ is obtained, it is compared to the set of codewords of M, using a decoding function $\delta(x^{\rho}, M)$. Thus, the final prediction is the class with the codeword that minimizes $\delta(x^{\rho}, M)$. Escalera et al. (2010a) proposed a problem-dependent decoding function (distance function that takes into account classifier performances) obtaining very satisfying results. Following this core idea, we use the Loss-Weighted decoding of Equation 3.1, where M_w is a matrix of weights and L is a loss function ($L(\theta) = \exp^{-\theta}$).

$$\delta_{LW}(x^s, i) = \sum_{j=1}^{n} M_w(i,j) L(y^i_j \cdot d^j(x^s)) \tag{3.1}$$

In Equation 3.1, M_w (weight matrix) corresponds to the product of cascade accuracy at each stage. Thus, each column i of M_w is assigned a weight w^i as,

$$w^i = \prod_{j=1}^{k} \frac{TP(d^i_j) + TN(d^i_j)}{TP(d^i_j) + FN(d^i_j) + FP(d^i_j) + TN(d^i_j)}, \tag{3.2}$$

for a cascade of classifiers of k stages, where d^i_j stands for the i-th cascade and stage j, $j \in [1, .., k]$, and TP, TN, FN, and FP computes the number of true positives, true negatives, false negatives and false positives, respectively. Finally, a body-like probability map $P^{bl} \in [0,1]^{l\times w}$, where l and w are the length and width of I, is build. This map contains, at each position P^{bl}_{ij}, the proportion of body part detections for each pixel over the total number of detections for the whole image. In other words, pixels belonging to the human body will show a higher body-like probability than the pixels belonging to the background. Examples of probability maps obtained from ECOC outputs are shown in Fig. 3.9(e) and 3.9(g), respectively. (see also step (c) in Fig. 3.5).

3.1.3.4 Binary GrabCut optimization for foreground mask extraction

GrabCut approach from Hernández-Vela et al. (2012b) has been widely used for interactive background/foreground extraction (binary segmentation). Formally, given a color image I, let us consider the array $z = (z_1, ..., z_q, ..., z_Q)$ of Q pixels where $z_i = (R_i, G_i, B_i)$, $i \in [1, ..., Q]$ in RGB space. The segmentation is defined as an array $\boldsymbol{\alpha} = (\alpha_1, ...\alpha_Q)$, $\alpha_i \in \{0, 1\}$, assigning a label to each pixel of the image indicating if it belongs to background or foreground. A trimap T is defined consisting of three regions: T_B, T_F and T_U, each one containing initial background, foreground, and uncertain pixels, respectively. Pixels

belonging to T_B and T_F are clamped as background and foreground respectively—which means GrabCut will not be able to modify these labels, whereas those belonging to T_U are actually the ones the algorithm will be able to label. Color information is introduced by GMMs. A full co-variance GMM of U components is defined for background pixels ($\alpha_i = 0$), and another one for foreground pixels ($\alpha_j = 1$), characterized as follows,

$$\boldsymbol{\theta} = \{\pi(\alpha, u), \mu(\alpha, u), \Sigma(\alpha, u), \alpha \in \{0, 1\}, u = 1..U\}, \tag{3.3}$$

being π the weights, μ the means and Σ the co-variance matrices of the model. We also consider the array $\mathbf{u} = \{u_1, ..., u_i, ...u_Q\}$, $u_i \in \{1, ...U\}$, $i \in [1, ..., Q]$ indicating the component of the background or foreground GMM (according to α_i) the pixel z_i belongs to. The energy function for segmentation E is then,

$$\mathbf{E}(\boldsymbol{\alpha}, \mathbf{u}, \boldsymbol{\theta}, \mathbf{z}) = \mathbf{U}(\boldsymbol{\alpha}, \mathbf{u}, \boldsymbol{\theta}, \mathbf{z}) + \lambda \mathbf{V}(\boldsymbol{\alpha}, \mathbf{z}), \tag{3.4}$$

where $\mathbf{U}$ is the likelihood potential based on the probabilities $p(\cdot)$ of the GMM,

$$\mathbf{U}(\boldsymbol{\alpha}, \mathbf{u}, \boldsymbol{\theta}, \mathbf{z}) = \sum_i -\log\, p(z_i|\alpha_i, u_i, \boldsymbol{\theta}) - \log\, \pi(\alpha_i, u_i), \tag{3.5}$$

and $\mathbf{V}$ is a regularizing prior assuming that segmented regions should be coherent in terms of color, taking into account a neighborhood $\mathcal{N}$ around each pixel,

$$\mathbf{V}(\boldsymbol{\alpha}, \mathbf{z}) = \gamma \sum_{\{m,q\} \in \mathcal{N}} [\alpha_q \neq \alpha_m] \exp\, (-\beta \|z_m - z_q\|^2), \tag{3.6}$$

where weight $\lambda \in \mathbb{R}^+$ specifies the relative importance of the boundary term against the unary term U.

With this energy minimization scheme and given the initial trimap T, the final segmentation is performed using a minimum cut algorithm. However, we propose to omit the classical semiautomatic trimap initialization by an automatic trimap assignment based on the human body probability map $P^{bl} \in [0, 1]^{l \times w}$. In this sense, depending on the probability of each pixel it will be assigned to a particular tag T_B, T_F and T_U.

3.1.3.5 Tree-structure body part learning without background

Once the binary person/background segmentation is performed utilizing GrabCut (mask shown in Fig. 3.5(e)), we apply a second procedure in order to split the person mask into a set of human limbs.

For this step, we define a new tree-structure classifier similar to the one described in Section 3.1.3.2 without including the background class c^7 shown in Fig. 3.6(a). An example of the tree-structure body part taking into account the set of 6 body limbs is shown in Fig. 3.7(a).

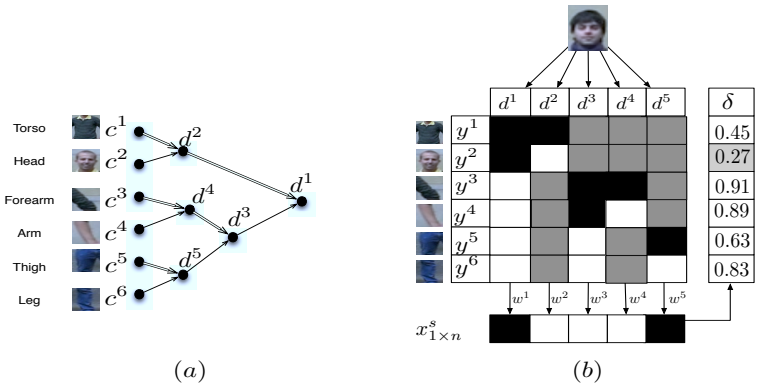


Figure 3.7: (a) tree-structure classifier of 6 body parts, (b) ECOC decoding step.

3.1.3.6 ECOC multi-limb detection fine-grained

In order to obtain an accurate detection of human limbs within the segmented user mask, we base on HOG descriptor from Dalal and Triggs (2005b) and SVM classifier which has shown to obtain robust results in human estimation scenarios as in Dalal and Triggs (2005b); Freund and Schapire (1995); Hernández-Vela et al. (2012b). We extract HOG features for the different body parts (previously normalized to dominant region orientation), so then, SVM classifiers are trained on that feature space, using a Generalized Gaussian RBF Kernel based on Chi-squared distance applied in Yang et al. (2009).

This stage follows a similar pipeline as the one described in Section 3.1.3.3. In this sense, each SVM classifier learns a binary partition of human limbs but without taking into account the background class. As shown in Fig. 3.6(b), we train $n = 6$ SVMs with different binary human-limb partitions.

At the ECOC decoding step, we also use the Loss-Weighted decoding function from Escalera et al. (2010a) shown in Equation 3.1 (an example is shown in Fig 3.7(b)). In this sense, for each RGB test image corresponding to the binary mask shown in Fig. 3.5(e), we adopt a sliding window approach and test each patch on our ECOC multi-limb recognition system. Then, based on the ECOC output we construct a set of limb-like probability maps. Each map P^c contains, at each position P^c_{ij}, the probability of pixel at the entry (i, j) of belonging to the body part class c, where $c \in \{1, 2, ..., 6\}$. This probability is computed as the proportion of detections at point (i, j) overall detection for class c. Examples of probability maps obtained from ECOC outputs are shown in Fig. 3.5(h). While Haar-like based on AdaBoost gave us a very accurate and fast initialization of human regions for binary user segmentation, in this second step, HOG-SVM is applied in a reduced region of the image, providing better estimates of human limb locations.

	Head	Torso	Arms	Forearms	Thighs	Legs	Background
Head	0	20	35	50	70	90	1
Torso	20	0	15	25	40	70	1
Arms	35	15	0	10	60	80	1
Forearms	50	25	10	0	30	60	1
Thighs	70	40	60	30	0	10	1
Legs	90	70	80	60	10	0	1
Background	1	1	1	1	1	1	1

Table 3.3: Prior cost between each pair of labels.

3.1.3.7 Alpha-beta swap Graph Cuts multi-limb segmentation

In our proposal, we base on Graph Cuts theory to tackle our human-limb segmentation problem as in Boykov and Funka-Lea (2006); Boykov and Kolmogorov (2003); Boykov et al. (2001); Hernández-Vela et al. (2012b); Rother et al. (2004a). Boykov et al. (2001) developed an algorithm, named α-β swap graph-cut, which can cope with the multi-label segmentation problem. The α-β swap graph-cut is an extension of binary graph cuts that performs an iterative procedure where each pair of labels (α_q, α_m), $\{m, q\} \in \{1, 2, ..., 6\}$, are segmented using GC. This procedure segment all α pixels from β pixels with GC and the algorithm will change the α-β combination at each iteration until convergence. However, to cope with the multi-label case, an extension of the minimization framework described in Section 3.1.3.4 is needed.

In this sense, $\alpha_i \in \{1, ..., c\}$ and an initial labeling $T \in \{T_1, ..., T_c\}$ is defined by an automatic trimap assignment based on the set of limb-like probability maps $P^c \in [0, 1]^{l \times w}$ defined in previous section. In addition, the coefficient that multiplies the exponential term in Equation 3.6, $[\alpha_q \neq \alpha_m]$, is changed to $\Omega(c_q, c_m)$, which penalizes relations between pixels z_q and z_m depending on their label assignations and a user-predefined pair-wise cost to each possible combination of labels,

$$\mathbf{V}(\mathbf{c}, \mathbf{z}) = \gamma \sum_{\{m,q\} \in \mathcal{N}} \Omega(c_q, c_m) \exp\left(-\beta \|z_m - z_q\|^2\right). \tag{3.7}$$

In concrete, in order to introduce prior costs between different labels, $\Omega(c_q, c_m)$ must fulfill some constraints related to spatial coherence between the different labels, taking into account the natural constraints of the human limbs (i.e., head must be closer to torso than legs, arms are nearer to forearms than head, etc.). In particular, we experimentally fixed the penalization function Ω as follows in Table 3.3:

3.1.4 Experimental results

In order to present the experimental results, we first discuss the data, experimental settings, methods and validation protocol.

3.1.4.1 Data

We use the proposed HuPBA $8k+$ dataset described in Section 3.1.2. We reduced the number of limbs from the 14 available in the dataset to 6, grouping those that are similar by symmetry (right-left) as arms, forearms, thighs, and legs. Thus, the set of limbs of our problem is head, torso, forearms, arms, thighs, and legs. Although labeled within the dataset, we did not include hands and feet in our multi-limb segmentation scheme. Finally, in order to train the limb classifiers, ground truth masks are used to normalize all limb regions per dominant orientation, and both Haar-like features and HOG descriptors are computed based on the aspect ratio of each region, being the descriptions scale invariant.

3.1.4.2 Methods and experimental settings

In this section we introduce the different methods compared for **binary segmentation**, **multi-limb segmentation** and **gesture recognition** tasks. In addition, the experimental settings for these methods are explained.

3.1.4.2.1 Binary segmentation

- **P.Detector+GbCut:** The well-known Person Detector of Dalal and Triggs (2005b) followed by GrabCut segmentation.
- **C.Class+GbCut:** The cascade of classifiers proposed by Viola and Jones (2001a), training one cascade of classifiers per limb and GrabCut segmentation.
- **ECOC+GbCut:** The proposed ECOC tree-structure body part classifier and automatic GrabCut segmentation.

3.1.4.2.2 Multi-limb segmentation

- **FMP:** This method was proposed by Yang and Ramanan (2011, 2013) and it is based on Flexible Mixtures-of-Parts (FMP). We compute the average of each set of mixtures for each limb and each pyramid level in order to obtain the probability maps for each limb category. In order to compute the probability map of the background category, we subtract 1 with the maximum probability of the set of limbs detection at the pixel location.

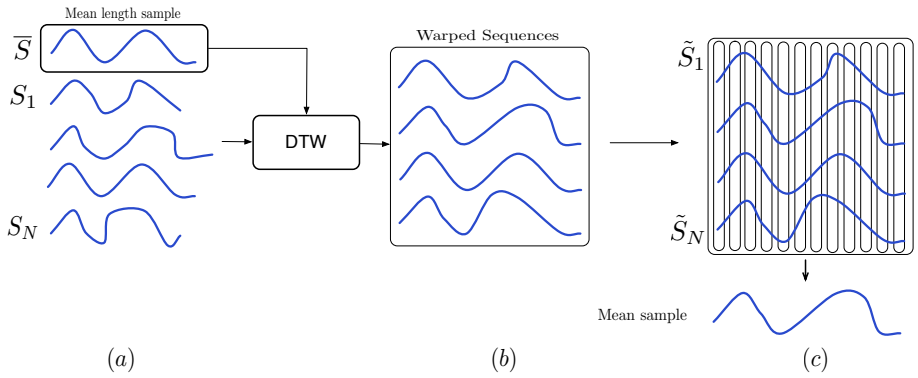


Figure 3.8: (a) Action samples and selected median length sample. (b) Aligned samples with same length . (c) Computation of the mean sample.

- **IPP:** This method is proposed by Ramanan (2006) and it is based on an Iterative Parsing Process (IPP). We use it to extract the limb-like probability maps followed by α-β swap graph-cut multi-limb segmentation. The background category is computed as shown in FMP method.
- **ECOC+GraphCut:** Our proposed human limb segmentation scheme shown in Fig. 3.5.

3.1.4.2.3 Gesture recognition

For the case of the gesture recognition task, our goal is to provide with a firm baseline of the recognition of the 11 actions categories labeled within the *HuPBA* $8K+$ dataset. In order to do it, we compare the performance of the following methodologies:

- **Dynamic Time Warping using a random sample**: We use the standard DTW algorithm to recognize the different actions categories in the dataset Sakoe et al. (1990). In order to compute the cost matrix for each of the gesture classes, we choose a sample of that category at random.
- **Dynamic Time Warping using the mean sample**: Following the trend in Hernández-Vela et al. (2013), to compute the cost matrix we form a mean sample of each one of the action classes. That is, we choose the sample of each category and align all samples with it. Then, once all samples from the same class are aligned (they have the same length) we compute the mean, an example is shown in Fig. 3.8. The cost-threshold for both DTW experiments was obtained by cross-validation on training data, using a leave-one-sequence-out procedure.
- **Hidden Markov Model**: We use the standard discrete HMM framework from Starner and Pentland (1997). Each HMM, was trained using the Baum-Welch al-

gorithm, and 3 states were experimentally set for the every action category, using a vocabulary of 10 symbols computed using K-means over the training data features. Final recognition is performed with temporal sliding windows of different wide sizes, based on the training samples length variability. The probability-threshold for the HMM experiment was obtained by cross-validation on training data, using a leave-one-sequence-out procedure.

3.1.4.2.4 Experimental settings

We used the standard Cascade of Classifiers based on AdaBoost from Viola and Jones (2001a), and we forced a 0.99 false positive rate and a maximum of 0.4 false alarm rate during 8 stages. In a preprocessing step, we resized al limb sample to a 32x32 pixels region for computational purposes. To detect limbs doing cascades of classifiers, we applied a sliding window approach with an initial patch size of 32x32 pixels up to 60x60 pixels. As a final part of the first stage, GrabCut was applied to obtain the binary segmentation where the initialization values of Foreground and background were provided to the GrabCut algorithm and tuned via cross-validation.

For the second stage, we set the following parameters for the HOG descriptor: 32x32 window size, 16x16 block size, 8x8 block stride, 8x8 cell size and 8 for several bins. Then, we trained SVMs with a Generalized Gaussian RBF kernel based on Chi-squared distance, (see Fig.(a) 3.7). The parameters of the kernel, C and γ were tuned via cross-validation. Finally, the model selection step was done via a leave-one-sequence-out CV. For multi-limb segmentation we used the GraphCut procedure where we tuned the λ parameter of GC, using CV setting an 8x8 neighboring grid.

3.1.4.3 Validation measurement

In order to evaluate the results for the three different tasks: binary segmentation, multi-label segmentation, and gesture recognition, we use the Jaccard Index ($J = \frac{A \bigcap B}{A \bigcup B}$) with the ground-truth.

3.1.4.4 Experimental Results

In this section we show both qualitative and quantitative results for the three different tasks: **binary segmentation**, **multi-label segmentation** and **gesture recognition**.

3.1.4.4.1 Binary segmentation

In Fig. 3.9 we can see an example of the person/background segmentation obtained by the compared methodologies. In particular, we can see in Fig. 3.9(d) how the segmentation

P.Detector+GbCut	C.Class+GbCut	ECOC+GbCut
49.60 ± 5.36	58.26 ± 4.24	$\mathbf{61.79 \pm 14.02}$

Table 3.4: Mean overlapping and standard deviation.

obtained by the Person Detector+GbCut method yields a poor result, segmenting dark regions of the image. Furthermore, when comparing Fig. 3.9(e) and 3.9(f), the improvement in the body-like probability map obtained by the ECOC+GbCut approach over the cascade class+GbCut method is significant.

In order to evaluate the performance of the compared methodologies, Table 3.4 shows the mean overlapping obtained on the whole dataset together with the standard deviation. From the results, one can see the ECOC+GbCut method outperforms the compared methodologies at least by a 5%. This improvement is the effect of two causes. The former is the Error-Correcting capabilities of the ECOC framework. The latter is the tree-structure definition of the coding matrix, which allows base classifiers to obtain accurate results.

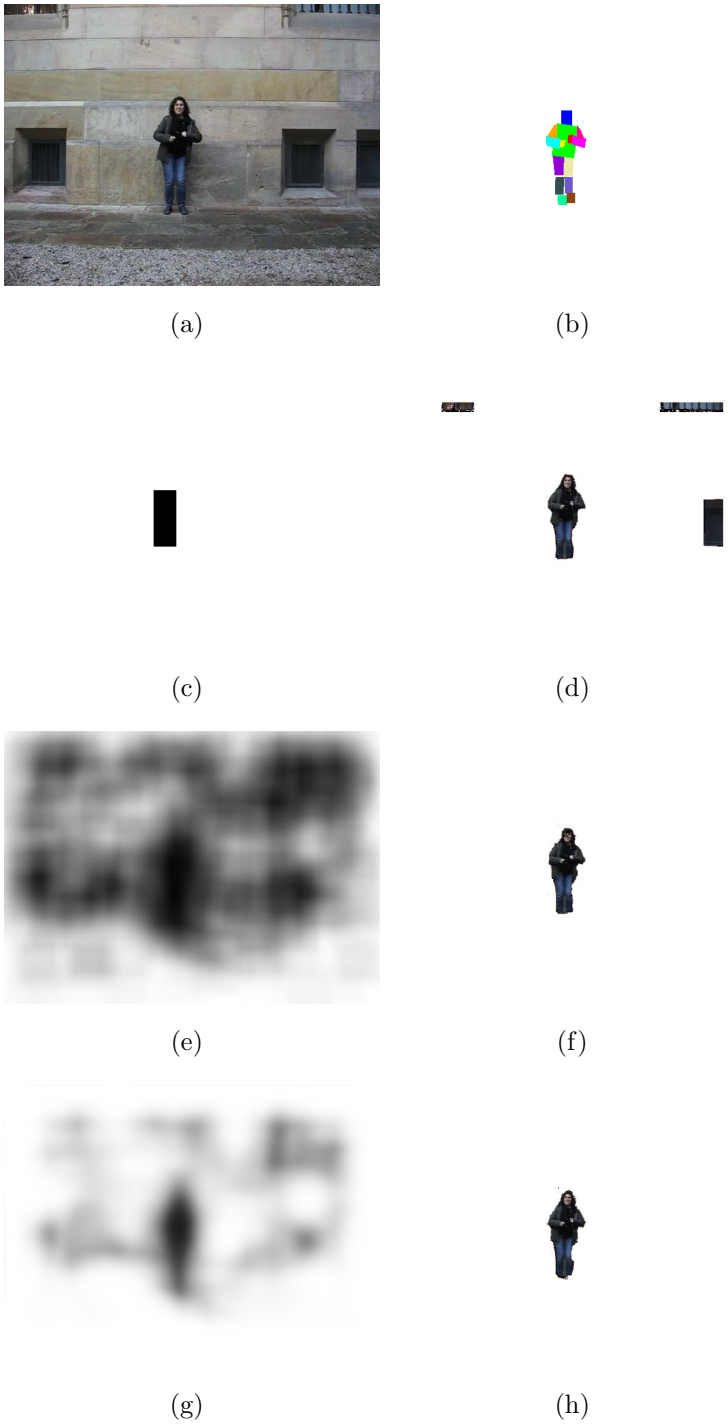
3.1.4.4.2 Multi-limb segmentation

Multi-limb segmentation, we show in Fig. 3.10 and Fig. 3.11 qualitative results. When comparing the qualitative results, we can see how the FMP method from Yang and Ramanan (2011, 2013) performs worse than its counterparts. Besides, one can see how IPP and our method obtain similar results.

Furthermore, we provide with quantitative results in terms of the Jaccard Index. In Fig. 3.12 we show the different overlapping performance obtained by the different methods, where each plot shows the overlapping for a particular limb. Besides, we analyze the overlapping performance as a function of a "Do not care" value that ranges from 0 to 4.

We use a "Do not care" value which provides a more flexible interpretation of the results. Consider the ground truth of a certain gesture category in a video sequence as a binary vector, which activates when a sample of such category is observed in the sequence. Then, the "Do not care" value is defined as the number of bits (frames) which are ignored at the limits of each one of the ground truth instances. Thus, by using this approach, we can compensate for the pessimistic overlap metric in situations when the detection has shifted some frames.

When analyzing quantitative results, we see how our method outperforms the compared methodologies most of the times. In particular, for the Head region, both methods obtain similar results, which is intuitive since the method used to detect the head is the well-known face detector. Finally, we see how FMP method is in all cases obtaining the worst performance.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3.9: (a) Original RGB image. (b) Multi-limb ground truth. (c) Probability map obtained by the Person Detector method. (d) Person/background segmentation of the Person Detector+GbCut approach. (e) Probability map yielded by the cascade class. method. (f) Person/background segmentation of the cascade class method. (g) Probability map obtained from the ECOC method. (h) RGB segmentation obtained by the ECOC+GbCut approach.

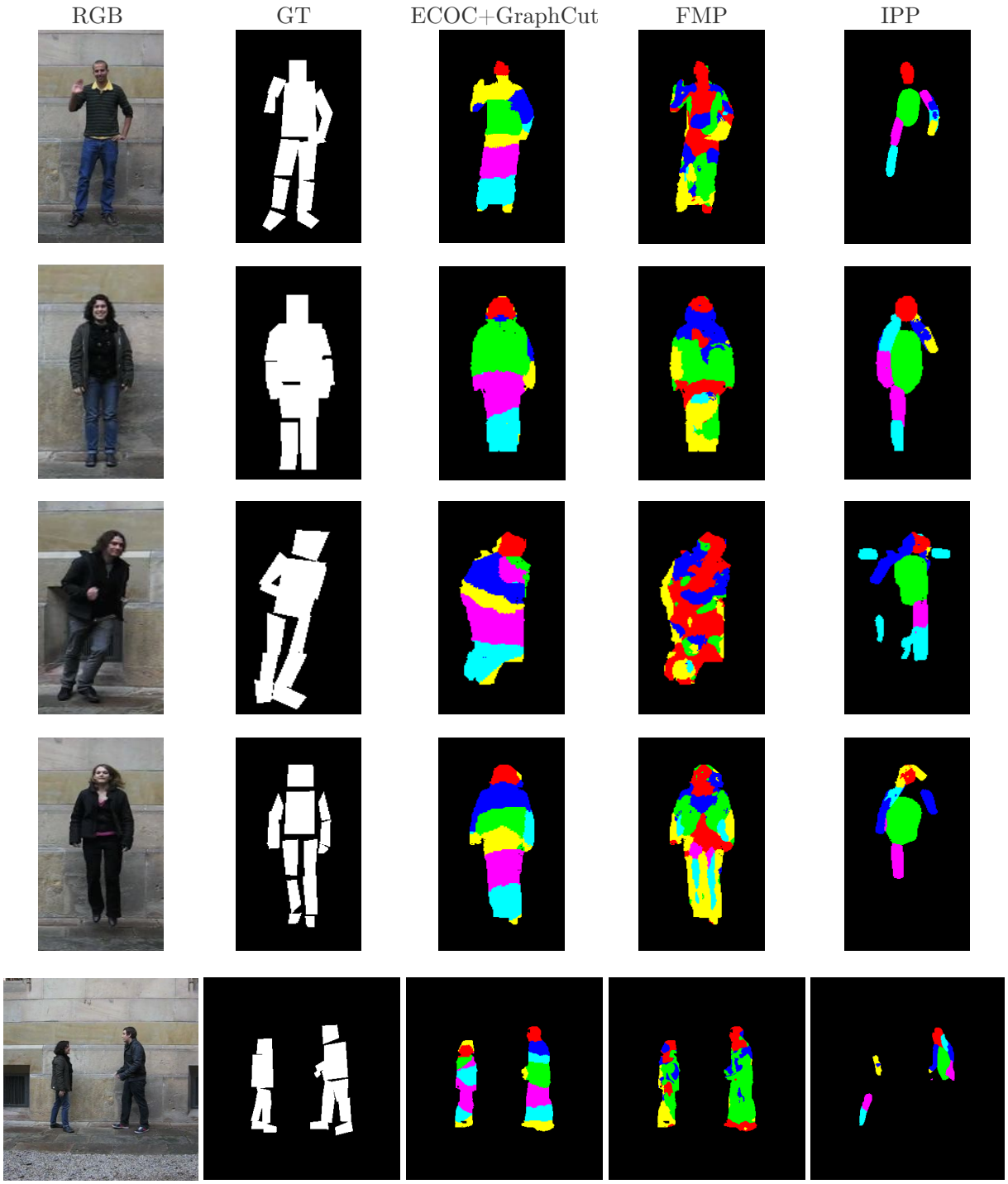


Figure 3.10: Multi-limb segmentation results for the three methods, for each sample, we also show the RGB image and the ground-truth (GT).

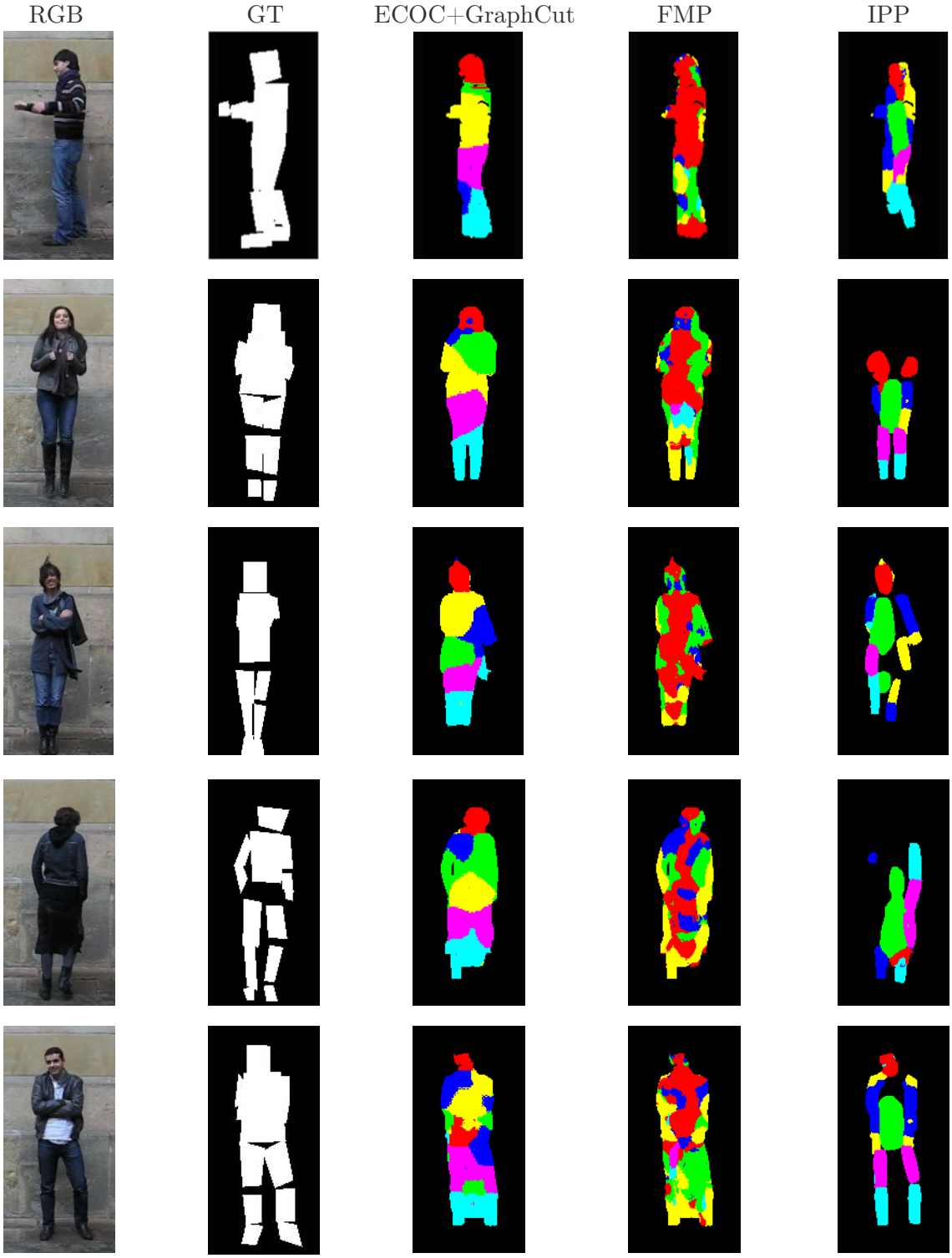


Figure 3.11: Multi-limb segmentation results for the three methods, for each sample, we also show the RGB image and the ground-truth (GT).

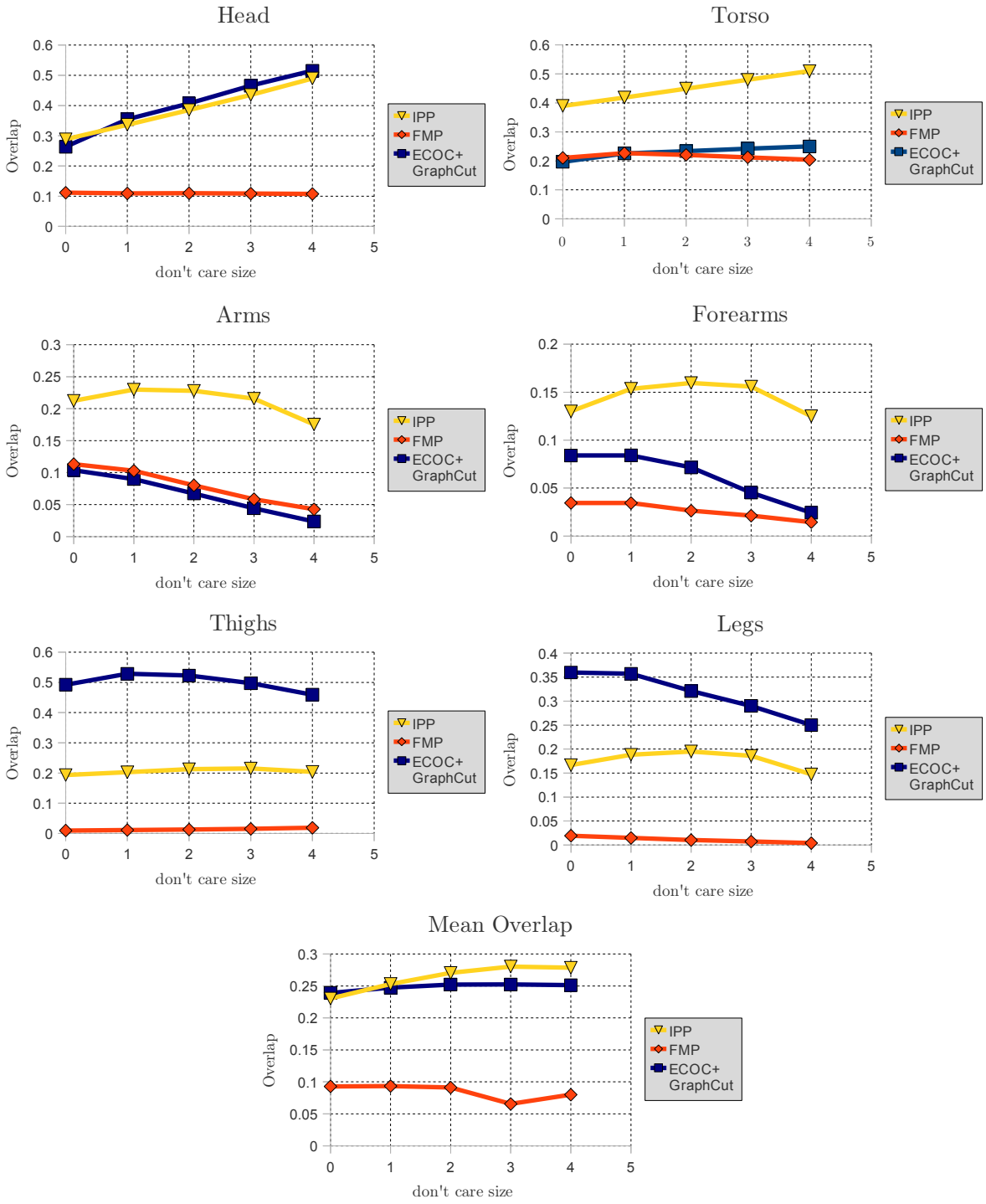


Figure 3.12: Overlap/Don't care size graph for each limb class and mean overlapping.

3.1.4.4.3 Gesture recognition

In this section, we show the quantitative results obtained by the different gesture recognition methods in terms of the Jaccard Index. Furthermore, to allow a more in-depth analysis of the proposed methodologies, in our evaluations we use a *Do not care* value which provides a more flexible interpretation of the results. Consider the ground truth of a particular action category in a video sequence as a binary vector, which activates when a sample of such category is observed in the sequence. Then, the *Do not care* value is defined as the number of bits (frames) which are ignored at the limits of each one of the ground truth instances. For further explanation of the algorithm see Bautista et al. (2015). Thus, by using this approach, we can compensate for the pessimistic overlap metric in situations when the detection has shifted some frames. The Jaccard Index as a function of the *Do not care* value for the 11 action categories and the mean Jaccard Index among action categories are shown in Fig. 3.13.

When analyzing quantitative results we see how the DTW Mean methods outperform for most action categories the standard DTW Random and HMM methods. Besides, when computing the mean Jaccard Index among all gesture categories the DTW Mean approach also ranks first, obtaining a mean Jaccard Index of 0.20. This good result is due to the use of information from all action samples which encodes the intra-class variability of the gesture categories. Finally, we can see how in every case the Hidden Markov Model is the worst performing method.

In the next section, we will see a particular approach for refining the limb-like probability maps by using Stacked Generalization Learning from Wolpert (1992). Following this approach, we will be able to train a second classifier, called meta-classifier which takes into account an extensive set of features that contains the likelihoods from previous classifiers.

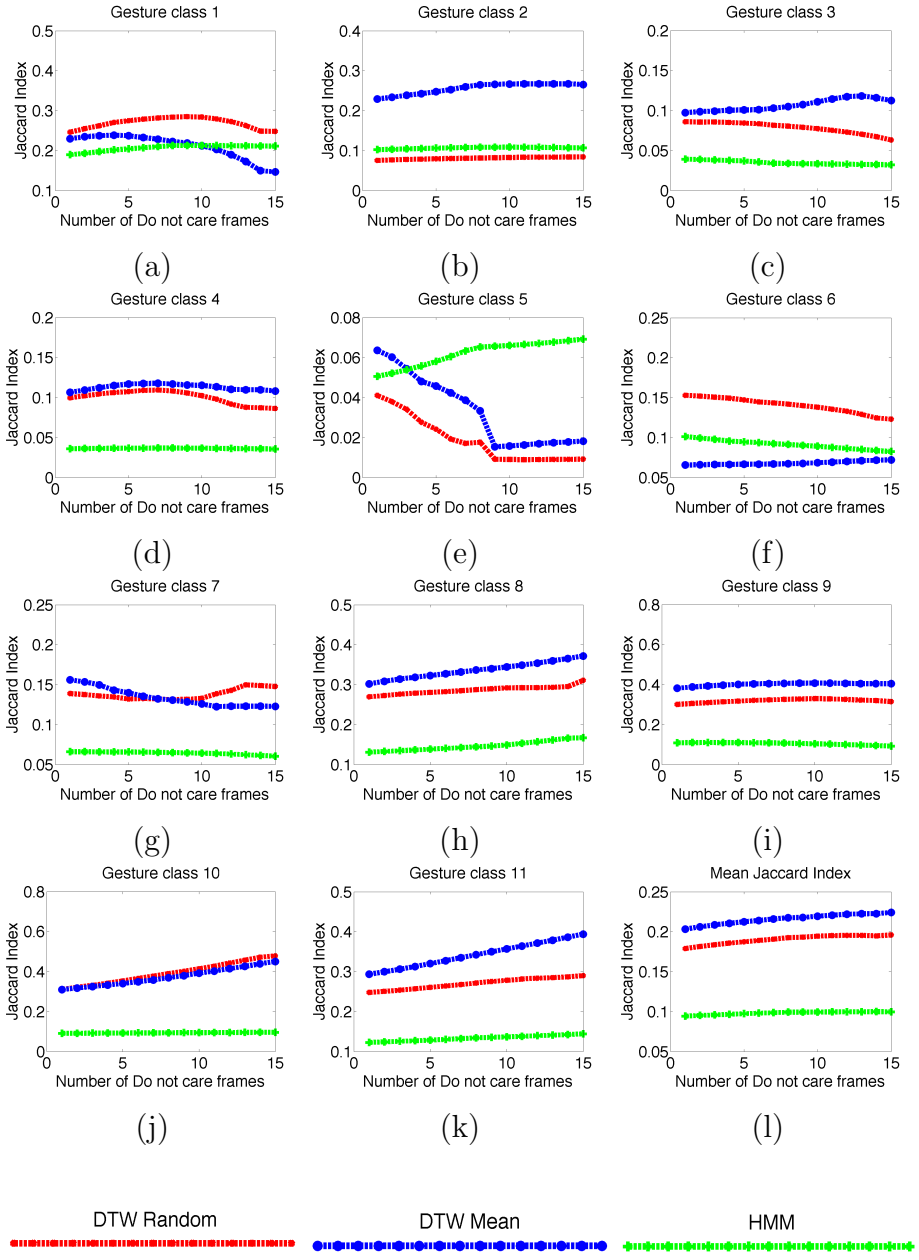


Figure 3.13: Jaccard Indexes for the different action categories from (a) to (k). (l) Shows the mean Jaccard Index among all action categories

3.2 Learning To Segment Humans By Stacking Their Body Parts

3.2.1 Introduction

Human segmentation in RGB images is a challenging task due to the high variability of the human body, which includes a wide range of human poses, lighting conditions, cluttering, clothes, appearance, background, point of view, number of human body limbs, etc. In this particular problem, the goal is to provide a complete segmentation of the person/people appearing in an image. In literature, human body segmentation is usually treated in a two-stage fashion. First, a human body part detection step is performed, obtaining a large set of candidate body parts. These parts are used as prior knowledge by segmentation/inference optimization algorithms in order to obtain the final human body segmentation.

In the first stage, that is the detection of body parts; weak classifiers are trained in order to obtain a soft prior of body parts (which are often noisy and unreliable). Most works in literature have used edge detectors, convolutions with filters, linear SVM classifiers, Adaboost or Cascading classifiers from Viola and Jones (2001b). For example, Ramanan (2006) used a tubular edge template as a detector and convolved it with an image defining locally maximal responses above a threshold as detections. Ramanan et al. (2005) used quadratic logistic regression on RGB features as the part detectors. Other works, have applied more robust part detectors such as SVM classifiers from Chakraborty et al. (2013); Hernández-Vela et al. (2012a,b) or AdaBoost in Pishchulin et al. (2013a) trained over HOG features from Dalal and Triggs (2005b). More recently, Dantone et al. (2013) used Random Forest as classifiers to learn body parts. Although recently robust classifiers have been used, part detectors still involve false-positive and false-negatives problems given the similarity nature among body parts and the presence of background artifacts. Therefore, a second stage is usually required in order to provide an accurate segmentation.

In the second stage, soft part detections are jointly optimized taking into account the nature of the human body. However, standard segmentation techniques (that is, region-growing, thresholding, edge detection, etc.) are not applicable in this context due to the large variability of environmental factors (i.e., lightning, clothing, cluttering, etc.) and the changing nature of body textures. In this sense, the most known models for the optimization/inference of soft part priors are the Poselets in Bourdev et al. (2010); Pishchulin et al. (2013a) and the Pictorial Structures in Andriluka et al. (2009); Felzenszwalb and Huttenlocher (2000); Sapp et al. (2010a) both of which optimize the initial soft body

part priors to obtain a more accurate estimation of the human pose, and provide with a multi-limb detection. Besides, some works in literature tackle the problem of human body segmentation (segmenting the full body as one class) obtaining satisfying results. For instance, Vineet et al. (2011) proposed to use Conditional Random Fields (CRF) based on body part detectors to obtain a complete person/background segmentation. Belief propagation, branch and bound or Graph Cut optimization are conventional approaches used to perform inference of the graphical models defined by the human body in Hernández-Vela et al. (2012a,b); Rother et al. (2004a). Finally, methods like structured SVM or a mixture of parts from Yang and Ramanan (2011); Yu and Joachims (2009) can be used in order to take profit of the contextual relations of body parts.

In this chapter, we present a novel two-stage human body segmentation method based on the discriminative Multi-Scale Stacked Sequential Learning (MSSL) framework from Gatta et al. (2011). Therefore, Ting and Witten (1997, 1999) studied firstly the type of generalizer that is more appropriate at a higher level model, also called meta-classifier or meta-regressor. Secondly, which type of features can be better used as input. It turned out that it is more convenient to use class probabilities than class predictions as input for the meta-model. Moreover, multi-response linear regression algorithm (MLR) gave the best results as a meta-model. Additionally, Stacked Learning has been applied for regression problems by Breiman (1996). Until now stacked sequential learning has been used in several domains, mainly in text sequences and time series from Carvalho and Cohen (2005); Dietterich (2002) showing significant computational and performance improvements when compared with other contextual inference methods such as CRF. Munoz et al. (2010) utilized several classifiers at different levels in order to define a hierarchical labeling strategy for semantic segmentation as a result of combining them per level in a stacking fashion. The work of Sun (2011) tackled a Natural Language Processing problem of Chinese word segmentation applying word, character, and local character classification and then stacked the segmented output sentences concerning efficiency and effectiveness. These and previous examples are approached in a 'cheap' manner to alternatives such as graphical models where there is a specific structured form. As another example, research on sociology and social media has been conducted by Dinakar et al. (2014), analyzing an online community supporting adolescents under duress by training different weak learners and combining their output in a stacked learning approach. Recently, the MSSL framework has also been successfully used on pixel-wise classification problems in Puertas et al. (2015). To the best of our knowledge, this is the first work that uses MSSL in order to find a context-aware feature set that encodes high order relations between body parts, which suffers non-rigid transformations, to obtain a robust human body segmentation. Fig. 3.14 shows the proposed human body segmentation approach. In the first stage of our method for human segmentation, a multi-class Error-Correcting Output Codes classifier (ECOC)

is trained to detect body parts and to produce a soft likelihood map for each body part. In the second stage, multi-scale decomposition of these maps and a neighborhood sampling is performed, resulting in a new set of features. The extended set of features encodes spatial, contextual and relational information among body parts. This extensive set is then fed to the second classifier of MSSL, in this case, a Random Forest binary classifier, which maps a multi-limb classification to a binary human classification problem. Finally, in order to obtain the resulting binary human segmentation, a post-processing step is performed through Graph Cuts optimization, which is applied to the output of the binary classifier.

3.2.2 Method

The proposed method for human body segmentation is based on the Multi-Scale Stacked Sequential Learning (MSSL) from Puertas et al. (2015) pipeline. Generalized Stacked Sequential Learning was proposed as a method for solving the main problems of sequential learning, namely: (a) how to capture and exploit sequential correlations; (b) how to represent and incorporate complex loss functions in contextual learning; (c) how to identify long-distance interactions; and (d) how to make sequential learning computationally efficient. Fig. 3.14 (a) shows the abstract blocks of the process[2]. Consider a training set consisting of data pairs $\{(x_i, y_i)\}$, where $x_i \in \mathcal{R}^n$ is a feature vector and $y_i \in \mathcal{Y}, \quad \mathcal{Y} = \{1, \ldots, K\}$ its class label. The first block consists of a classifier $H_1(x)$ trained with the input data set. The output results in a set of predicted labels or confidence values Y'. The next block in the pipeline defines the policy for taking into account the context and long-range interactions. It is composed of two steps: first, a multi-resolution decomposition models the relationship among nearby locations, and second, a neighborhood sampling proportional to the resolution scale defines the support lattice. This last step allows for modeling the interaction range. This block is represented by the function $z = J(x, \rho, \theta) : \mathcal{R} \rightarrow \mathcal{R}^w$, characterized by the interaction range θ in a neighborhood ρ. The last step of the algorithm creates an extended data set by adding to the original data the new set of features resulting from the sampling of the multi-resolution confidence maps which is the input of a second classifier $H_2(x)$.

3.2.2.1 Stage One: Body Parts Soft Detection

In this chapter, the first stage detector $H_1(x)$ in the MSSL pipeline is based on the soft body parts detectors defined in Sánchez et al. (2013). The work of Sánchez et al. (2013) is based on an ECOC ensemble of cascades of AdaBoost classifiers. Each of the

[2]The original formulation of MSSL also includes the input vector X as an additional feature in the extended set X'.

cascades focuses on a subset of body parts described using Haar-like features where regions have been previously moved towards main orientation to make the recognition rotation invariant. Although any other part detector the technique could be used in the first stage of our process, we also choose the same methodology. ECOC has shown to be a powerful and the general framework that allows the inclusion of any base classifier, involving error-correction capabilities and allowing to reduce the bias and variance errors of the ensemble as in Dietterich and Bakiri (1994); Escalera et al. (2008). As a case study, although any classifier can be included in the ECOC framework, here we considered as a base learner also the same ensemble of cascades given its fast computation.

Because of its properties, a cascade of classifiers is usually trained to split one visual object from the rest of the possible objects of an image. This means that the cascade of classifiers learns to detect a particular object (body part in our case), ignoring all other objects (all other body parts). However, somebody parts have a similar appearance, that is, legs and arms, and thus, it makes sense to group them in the same visual category. Because of this, we learn a set of cascades of classifiers where a subset of limbs are included in the positive set of one cascade, and the remaining limbs are included as negative instances together with background images in the negative set of the cascade. In this sense, classifier H_1 is learned by different grouping cascades of classifiers in a tree-structure way and combining them in an Error-Correcting Output Codes (ECOC) framework as Escalera et al. (2010b). Then, H_1 outputs correspond to a multi-limb classification prediction.

An example of the body part tree-structure defined taking into account the nature of human body parts is shown in Fig. 3.15(a). Notice that classes with similar visual appearance (that is, upper-arm and lower-arm) are grouped in the same meta-class in most dichotomies. Besides, dichotomies that deal with difficult problems (that is, d^5) are focused only on the difficult classes, without taking into account all other body parts. In this case, class c^7 denotes the background.

In the ECOC framework, given a set of K classes (body parts) to be learned, m different bi-partitions (groups of classes or dichotomies) are formed, and n binary problems over the partitions are trained as in Bautista et al. (2012a). As a result, a codeword of length n is obtained for each class, where each position (bit) of the code corresponds to a response of a given classifier d (coded by $+1$ or -1 according to their class set membership, alternatively, 0 if a particular class is not considered for a given classifier). Arranging the codewords as rows of a matrix, we define a *coding matrix* M, where $M \in \{-1, 0, +1\}^{K \times n}$. During the *decoding* (or testing) process, applying the n binary classifiers, a code c is obtained for each data sample x in the test set. This code is compared to the base codewords ($y^i, i \in \{1, .., K\}$[3]) of each a class defined in the matrix M, and the data

[3]Observe that we are overloading the notation of y so that y^i corresponds to the codeword of the

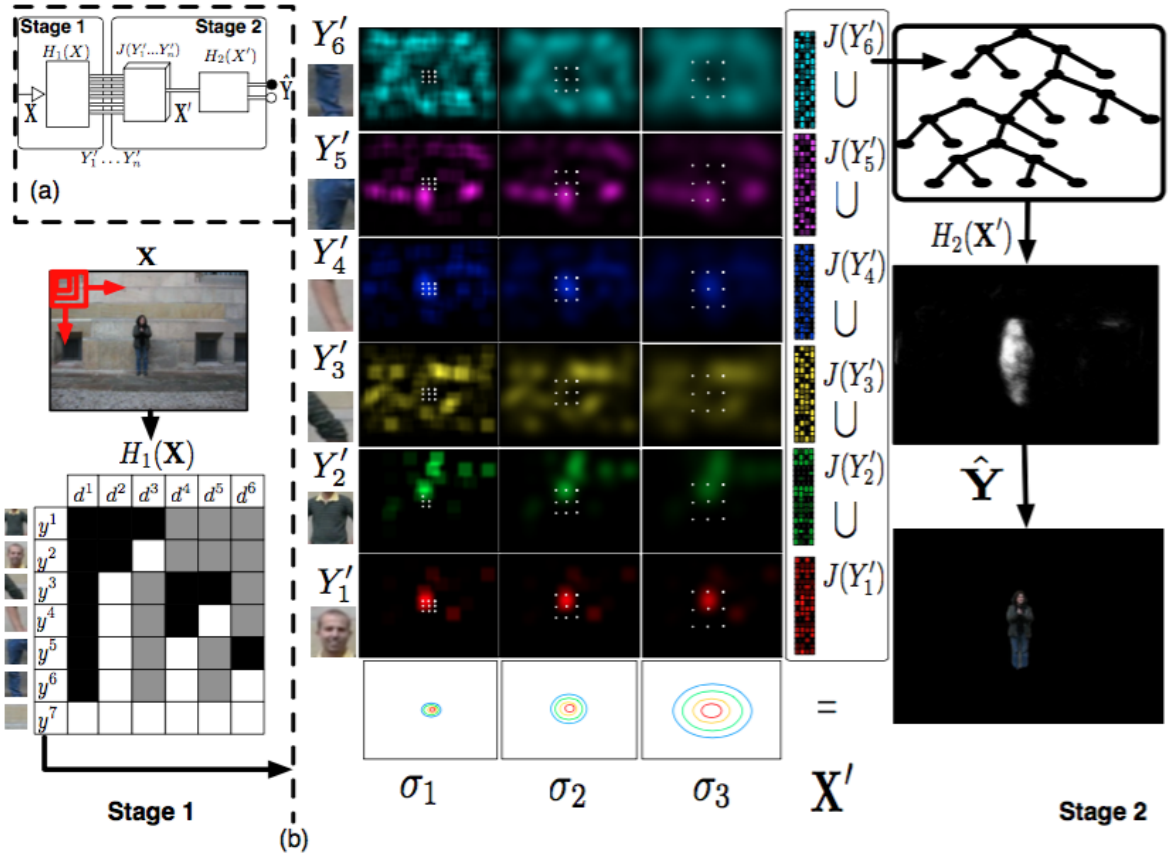


Figure 3.14: Method overview. (a) Abstract pipeline of the proposed MSSL method where the outputs Y'_i of the first multi-class classifier $H_1(x)$ are fed to the multi-scale decomposition and sampling function $J(x)$ and then used to train the second stacked classifier $H_2(x)$ which provides a binary output $\hat{\mathbf{Y}}$. (b) Detailed pipeline for the MSSL approach used in the human segmentation context where $H_1(x)$ is a multi-class classifier that takes a vector $\mathbf{X}$ of images from a dataset. As a result, a set of likelihood maps $Y'_1 \dots Y'_n$ for each part is produced. Then a multi-scale decomposition with a neighborhood sampling function $J(x)$ is applied. The output $\mathbf{X}'$ produced is taken as the input of the second classifier $H_2(x)$, which produces the final likelihood map $\hat{\mathbf{Y}}$, showing for each point the confidence of belonging to human body class.

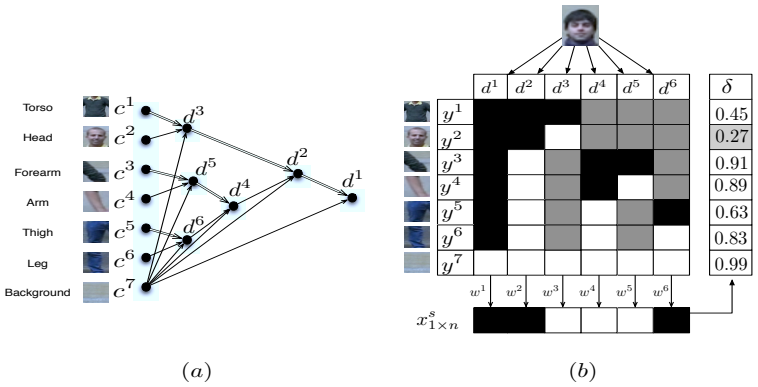


Figure 3.15: (a) Tree-structure classifier of body parts, where nodes represent the defined dichotomies. Notice that the single or double lines indicate the meta-class defined. (b) ECOC decoding step, in which a head sample is classified. The coding matrix codifies the tree-structure of (a), where black and white positions are codified as $+1$ and -1, respectively. c, d, y, w, X, and δ correspond to a class category, a dichotomy, a class codeword, a dichotomy weight, a test codeword, and a decoding function, respectively.

sample is assigned to the class with the *closest* codeword as in Escalera et al. (2010b).

We use the problem dependent coding matrix defined in Sánchez et al. (2013) in order to allow the inclusion of cascade of classifiers and learn the body parts. In particular, each dichotomy is obtained from the body part tree-structure. Fig. 3.15(b) shows the coding matrix codification of the tree-structure in Fig. 3.15(a).

In the ECOC *decoding* step an image is processed using a sliding windowing approach. Each image patch x is described and tested. In our case, each patch is first rotated by main gradient orientation and tested using the ECOC ensemble with Haar-like features and cascade of the classifier. In this sense, each classifier d outputs a prediction whether x belongs to one of the two previously learned meta-classes. Once the set of predictions $c \in \{+1, -1\}^{1\times n}$ is obtained, it is compared to the set of codewords of the classes y^i from M, using a decoding function $\delta(c, y^i)$ and the final prediction is the class with the codeword with minimum decoding, that is, $\arg\min_i \delta(c, y^i)$. As a decoding function, we use the Loss-Weighted approach with linear loss function defined in Escalera et al. (2010b). Then, a body-like probability map is built. This map contains, at each position the proportion of body part detections for each pixel over the total number of detections for the whole image. In other words, pixels belonging to the human body will show a higher body-like probability that the pixels belonging to the background. Additionally, we also construct a set of limb-like probability maps. Each map contains at each position (i, j) the probability of pixel at the entry (i, j) of belonging to the body part class. This

matrix associated with class i, that is, it is the i-th row of the matrix, $M(i, :)$.

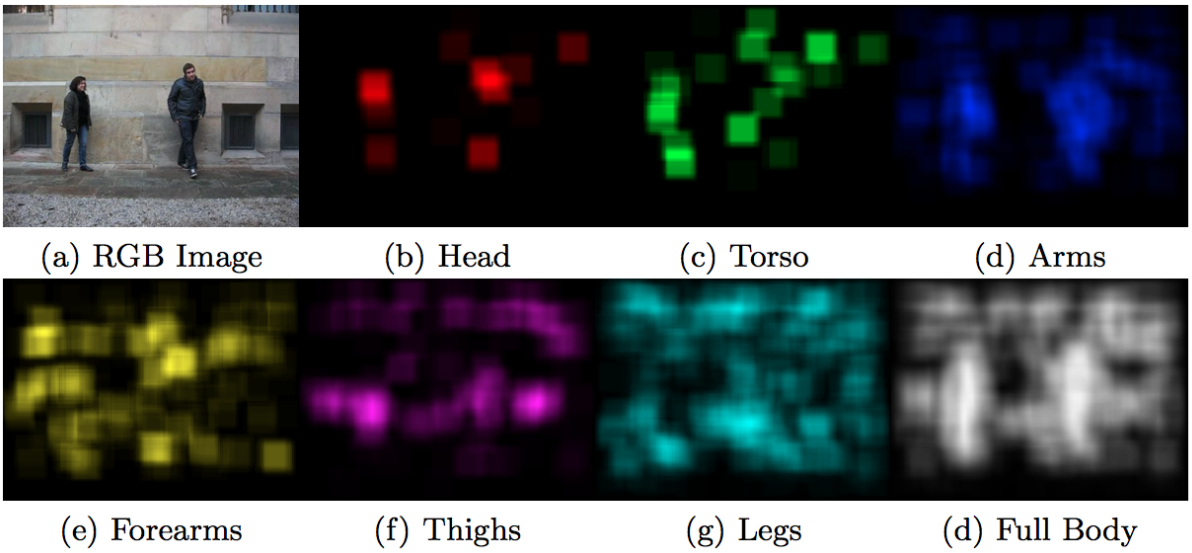


Figure 3.16: Limb-like probability maps for the set of 6 limbs and body-like probability map. Image (a) shows the original RGB image. Images from (b) to (g) illustrate the limb-like probability maps and (h) shows the union of these maps.

probability is computed as the proportion of detections at point (i, j) overall detection for that class. Examples of probability maps obtained from ECOC outputs are shown in Fig. 3.16, which represents the $H_1(x)$ outputs $Y'_1 \dots Y'_n$ defined in Fig. 3.14 (a).

3.2.2.2 Stage Two: Fusing Limb Likelihood Maps Using MSSL

The goal of this stage is to fuse all partial body parts into a full human body likelihood map (see Fig. 3.14 (b) the second stage). The input data for the neighborhood modeling function $J(x)$ are the body parts likelihood maps obtained in the first stage $(Y'_1 \dots Y'_n)$. In the first step of the modeling, a set of different Gaussian filters is applied on each map. All these multi-resolution decompositions give information about the influence of each body part at different scales along with space. Then, an 8-neighbor sampling is performed for each pixel with a sampling distance proportional to its decomposition scale. This allows taking into account the different limbs influence and their context. The extended set X' is formed by stacking all the resulting sampling at each scale for each limb likelihood map (see the extended feature set X' in Fig. 3.14(b)). As a result, X' will have dimensionality equals the number of sampling multiplied by the number of scales and the number of body parts. In our experiments, we use eight neighbor sampling, three scales, and six body parts. Notice that contrary to the MSSL traditional framework, we do not feed the second classifier H_2 with both the original X and extended X' features, and only the extended set X' is provided. In this sense, the goal of H_2 is to learn spatial relations

among body parts based on the confidences produced by the first classifier. As a result, the second classifier provides a likelihood of the membership of an image pixel to the class 'person'. Thus, the multiple spatial relations of body parts (obtained as a multi-class classifier in H_1), are labelled as a two-class problem (*person* vs *no person*) and trained by H_2. Consequently, the label set associated with the extended training data X' corresponds to the union of the ground truths of all human body parts. Although within our method, any binary classifier can be considered for H_2, we use a Random Forest classifier to train 50 random trees that focus on different configurations of the data features. This strategy has shown robust results for human body segmentation in multi-modal data as in Shotton et al. (2013). Fig. 3.17 shows a comparison between the union of the likelihood maps obtained by the first classifier and the final likelihoods obtained after the second stage. We can see that a naive fusion of the limb likelihoods produces noisy outputs in many body parts. The last column shows how the second stage detects the human body using the same data. For instance, Fig. 3.17 (f) shows how it works well also when two bodies are a close one to others, splitting them accurately, preserving the poses. Notice that in Fig. 3.17 (f) there is a different to zero both silhouettes, existence of handshaking. Finally in Fig. 3.17 (c) we can see how the foreground person is highlighted in the likelihood map, while in previous stage (Fig. 3.17 (b)) it was completely missed. This shows that the second stage can restore body objects at different scales. Finally, the output likelihood maps obtained after this stage are used as input of a post-process based on graph-cut to obtain final segmentation.

3.2.3 Experimental Results

Before present the experimental results, we first discuss the data, experimental settings, methods, and validation protocol.

3.2.3.1 Dataset

We used *HuPBA 8k+ dataset* described in Sánchez et al. (2015). This dataset contains more than 8000 labeled images at pixel precision, including more than 120000 manually labeled samples of 14 different limbs. The images are obtained from 9 videos (RGB sequences), and a total of 14 different actors appear in those 9 sequences. In concrete, each sequence has the main actor (9 in total) which during the sequence interacts with secondary actors portraying a wide range of poses. For our experiments, we reduced the number of limbs from the 14 available in the dataset to 6, grouping those that are similar by symmetry (right-left) as arms, forearms, thighs, and legs. Thus, the set of limbs of our problem is composed by: *head*, *torso*, *forearms*, *arms*, *thighs* and *legs*. Although labeled within the dataset, we did not include hands and feet in our segmentation scheme. In

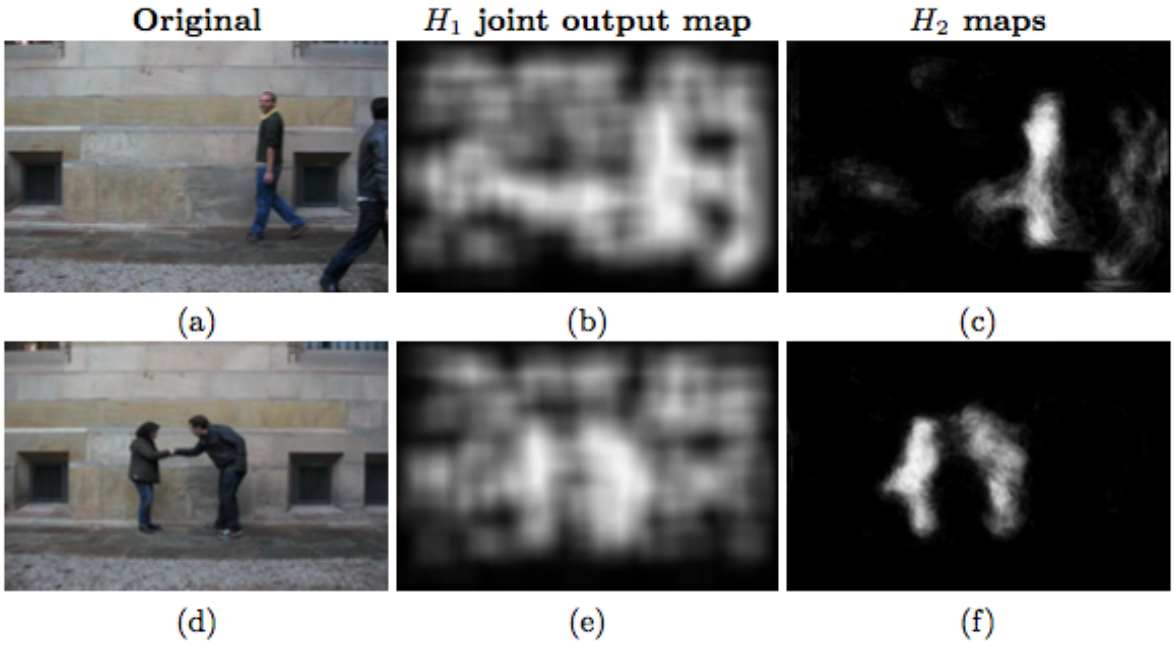


Figure 3.17: Comparative between H_1 and H_2 output. First column are the original images. Second column are H_2 output likelihood maps. Last column are the union of all likelihood map of body parts

Fig. 3.18 some samples of the *HuPBA* $8k+$ dataset are shown.

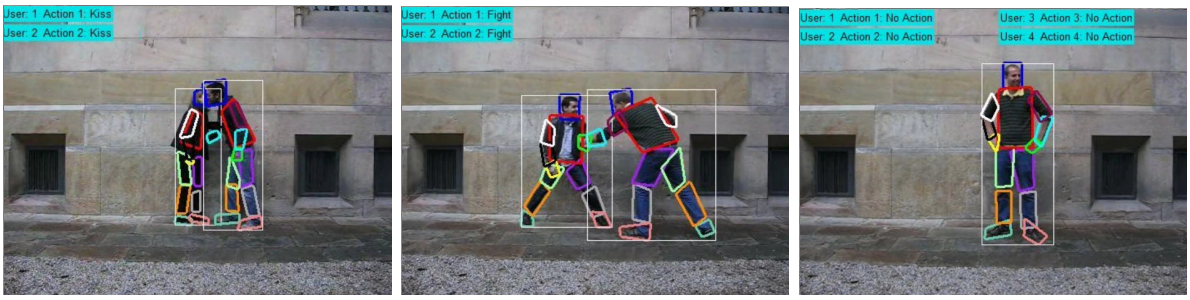


Figure 3.18: Different samples of the HuPBA $8k+$ dataset.

3.2.3.2 Methods

We compare the following methods for Human Segmentation: **Soft Body Parts (SBP) detectors + MSSL + Graphcut**. The proposed method, where the body like confidence map obtained by each body part soft detector is learned employing MSSL, and the output is then fed to a GraphCut optimization to obtain the final segmentation. **SBP detectors + MSSL + GMM-Graphcut**. Variation of the proposed method, where the

final GraphCut optimization also learns a GMM color model to obtain the final segmentation as in the GrabCut model Rother et al. (2004b). **SBP detectors + GraphCut**. In this method, the body like confidence map obtained by aggregating all body parts soft detectors outputs is fed to a GraphCut optimization to obtain the final segmentation. **SBP detectors + GMM-GraphCut**. We also use the GMM color modeling variant in the comparison.

3.2.3.3 Settings and validation protocol

In a preprocessing step, we resized all limb samples to a 32×32 pixels region. Regions are first rotated by main gradient orientation. In the first stage, we used the standard Cascade of Classifiers based on AdaBoost and Haar-like features from Viola and Jones (2001b) as our body part multi-class classifier H_1. As model parameters, we forced a 0.99 false positive rate and maximum of 0.4 false alarm rate during 8 stages. To detect limbs with trained cascades of classifiers, we applied a sliding window approach with an initial patch size of 32×32 pixels up to 60×60 pixels. As a result of this stage, we obtained 6 likelihood maps for each image. In the second stage, we performed 3-scale Gaussian decomposition with $\sigma \in [8, 16, 32]$ for each body part. Then, we generated an extensive set selecting for each pixel its 8-neighbors with σ displacement. From this extensive set, a sampling of 1500 selected points formed the input examples for the second classifier. As the second classifier, we used a Random Forest with 50 decision trees. Finally, in a post-processing stage, binary Graph Cuts with a GMM color modeling (we experimentally set 3 components) were applied to obtain the binary segmentation where the initialization seeds of foreground and background were tuned via cross-validation. For the binary Graph Cuts without a GMM color modeling we directly fed the body likelihood map to the optimization method. In order to assess our results, we used 9-fold cross-validation, where each fold correspond to images of the main actor sequence. As results the measurement we used the Jaccard Index of overlapping ($J = \frac{A \bigcap B}{A \bigcup B}$) where A is the ground-truth also, B is the corresponding prediction.

3.2.3.4 Quantitative Results

In Table 3.5 we show overlapping results for the *HuPBA* $8K+$ dataset. Specifically, we show the mean overlapping value obtained by the compared methods on 9 folds of the *HuPBA* $8k+$ dataset. We can see how our MSSL proposal consistently obtains a higher overlapping value on every fold.

Notice that MSSL proposal outperforms in the SBP+GC method in all folds (by at least a 3% difference), which is the state-of-the-art method for human segmentation in the HuPBA $8k+$ dataset from Sánchez et al. (2013).

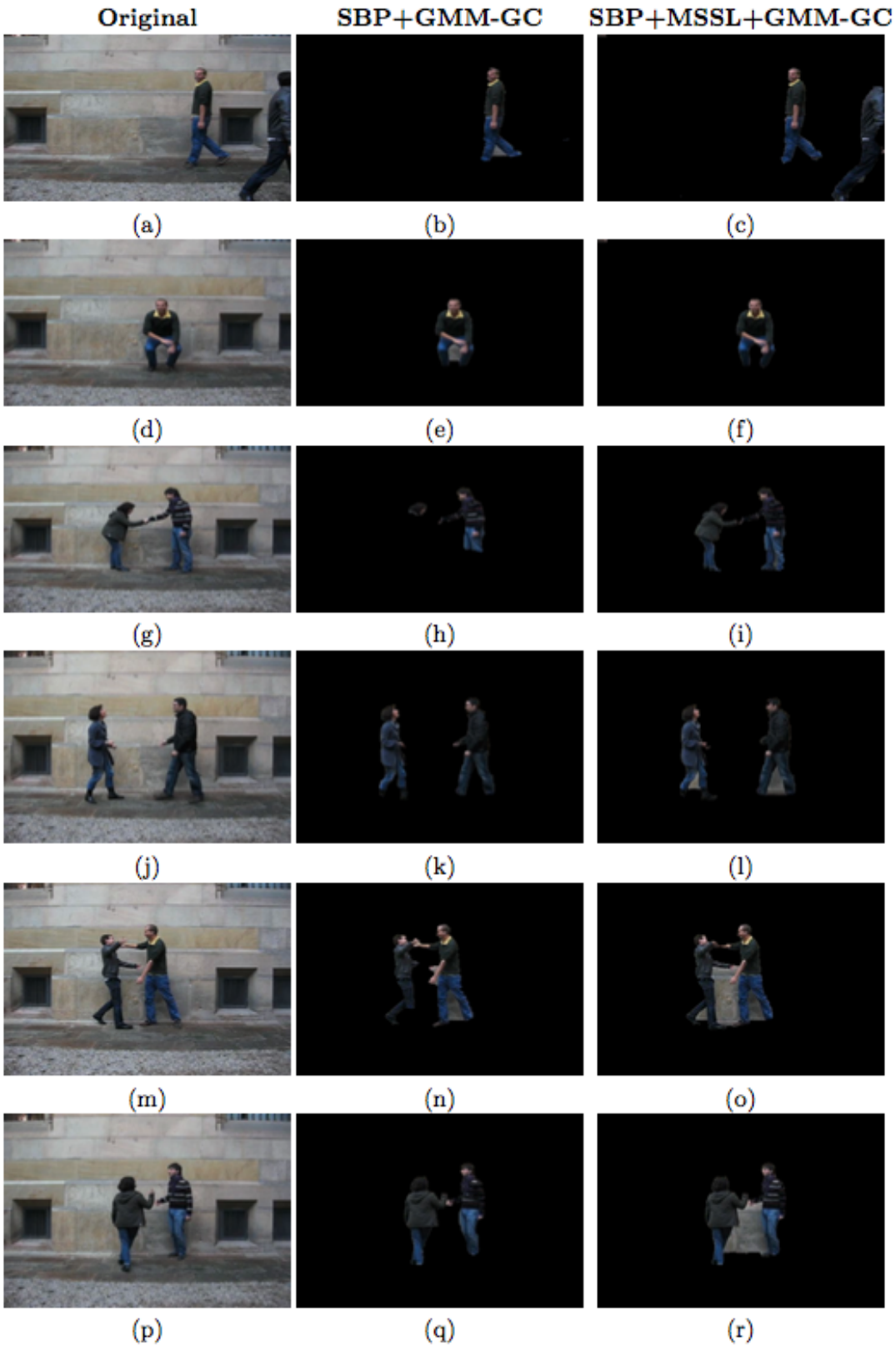


Figure 3.19: Samples of the segmentation results obtained by the compared approaches.

	GMM-GC		GC	
	MSSL	**Soft Detect.**	**MSSL**	**Soft Detect.**
Fold	**Overlap**	**Overlap**	**Overlap**	**Overlap**
1	**62.35**	60.35	**63.16**	60.53
2	**67.77**	63.72	**67.28**	63.75
3	**62.22**	60.72	**61.76**	60.67
4	**58.53**	55.69	**58.28**	55.42
5	**55.79**	51.60	**55.21**	51.53
6	**62.58**	56.56	**62.33**	55.83
7	**63.08**	60.67	**62.79**	60.62
8	**67.37**	64.84	**67.41**	65.41
9	**64.95**	59.83	**64.21**	59.90
Mean	**62,73**	59,33	**62,49**	59,29

Table 3.5: Overlapping results over the 9 folds of the *HupBA8K+* dataset for the proposed MSSL method and the Soft detectors post-processing their outputs with the Graph-Cuts method and GMM Graph-Cuts method.

3.2.3.5 Qualitative Results

In Fig. 3.19 some qualitative results of the compared methodologies for human segmentation are shown. It can be observed how in general SBP+MSSL+GMM-GC obtains a better segmentation of the human body than the SBP + GMM-GC method. This improvement is due to the contextual body part information encoded in the extended feature set. In particular, this performance difference is clearly visible in Fig. 3.19(f) where the human pose is completely extracted from the background. We also observe how the proposed method can detect a significative number of body parts at different scales. This is clearly appreciated in Fig. 3.19(c), where persons at different scales are segmented, while in Fig. 3.19(b) the SBP+GMM-GC fails to segment the rightmost person. Furthermore, Fig. 3.19(i) shows how the proposed method can recover the whole body pose by stacking all body parts, while in Fig. 3.19(h) the SBP+GMM-GC method just detected the head of the left most user. In this pair of images also we can see how our method can discriminate the different people appearing in an image, segmenting as background the interspace between them. Although, it may cause some loss, especially in the thinner body parts, like happens with the extended arm. Due to space restrictions, a table with more examples of segmentation results can be found in the supplementary material. Regards the dataset used, it is important to remark the large amount of segmented bodies (more than 10.000) and their high variability in terms of pose (performing different activities

and interactions with different people), size and clothes. The scale variations are learned by H_2 through spatial relationships of body parts. In addition, although background is maintained across the data, H_2 is trained over the soft predictions from H_1 (see the large number of false positive predictions shown in Fig. 3.16), and our method considerably improves those person confidence maps, as shown in Fig. 3.17.

At this point, we have seen different methodologies to tackle human multi-limb or full-body segmentation by using traditional state-of-the-art techniques where the pipeline is divided into several stages. That is, data preparation, feature extraction, model training, and predictions are done separately. The next chapter is focused on an alternative way of approaching human pose and segmentation problems through using Deep Convolutional Neural Networks (DCNN) and Multi-Task Learning (MTL) paradigm. The former is a hierarchical feature extractor and a learnable model framework which extracts features from samples and optimizes a decision function at the same time. Thus, some of these stages are done altogether in the same framework and not separately. The latter is a well-known paradigm in machine learning that has not been yet exploited in Deep Learning, and it aims to divide a problem or task into subtasks in order to ease the learning procedure and to make each subtask help each other.

3.3 Conclusions

The main conclusions of the two works presented in this chapter are summarized in the following sections:

3.3.1 HuPBA 8k+: Dataset and ECOC-GraphCut based Segmentation of Human Limbs

We defined in this chapter a novel dataset introduced in the Chalearn ECCV'14 Escalera et al. (2014) that consists of around 8000 images of human poses and annotated 14 body-limbs to tackle either the human multi-limb segmentation or the human pose estimation problem.

Moreover, we introduced a two-stage approach for human multi-limb segmentation that reduced in each stage the multi-limb search space. First, a set of AdaaBost cascades with Haar-like features were trained on top of an ECOC framework for human binary classification. Then, once the human body was obtained, a set of SVM's with HOG features was trained on top of an ECOC in order to get the limb-like probability maps. Finally, these maps were used to initialize GraphCuts to obtain the final segmentation. The current approach was compared over two state-of-the-art pose estimation approaches obtaining noticeably higher performance.

3.3.2 Learning to segment humans by stacking their body parts

In this chapter, we focused on the Stacked Generalization Learning approach which is a type of ensemble learning method. Concretely, we made use of a two-stage scheme based on the MSSL framework for human body segmentation. In the first stage, a pipeline consisting in AdaBoost cascades with Haar-like features initialized a set of soft limb-like probability maps to be stacked in a second classifier to infer finer detections. This second classifier learned co-dependencies among features and spatial relationships, and it was tested over state-of-the-art methods reaching accurate results.

Chapter 4

Block II

In this chapter, we make use of multi-task learning paradigm to approach multiple tasks where human body part segmentation is among them. Besides, we study deep learning methods to fuse the multiples tasks jointly in one model and not separately. Finally, we analyze in detail which tasks benefits or helps each other in order to validate the performance both per task and globally.

4.1 Multi-task human analysis in still images: 2D/3D pose, depth map, and multi-part segmentation

4.1.1 Introduction

Nowadays large amounts of annotated (or weakly annotated) data are publicly available for the automatic analysis of humans. Lin et al. (2014) collect a vast richly-annotated data for image classification, object localization, semantic/instance segmentation which person category is defined plus animals and objects. Besides, Everingham et al. (2015) publish one of the first large datasets including person among other classes for detection, classification and segmentation tasks. Moreover, Varol et al. (2017) release a large-scale dataset consisting of realistic synthetic data captured by MoCap and covering several cues: body depth maps, 2D/3D coordinates, body part segmentation, optical flow and surface normal. The dataset of Liang et al. (2018) is made up of multiple people which cues are 2D coordinates, body part segmentation and clothes parsing. Related tasks include several 2D pose estimation, body part segmentation, clothes parsing. Andriluka et al. (2014) release a larger dataset than previous with a high degree of variability in the human pose, viewpoint and covering a wide range of daily activities. Gong et al. (2017) make public a human pose and semantic part labels which range from the viewpoint, occlusions, and background complexity. Lassner et al. (2017) extend previous smaller

datasets with body parts segmentation and more joint coordinates by fitting a gender-neutral body model into the images. Nie et al. (2017) define a multiple person pose estimation by leveraging centroids embeddings in a dense joint regression deep neural network. Similarly, Liang et al. (2018) contribute state-of-the-art with multiple people in the images and diversify segmentation in clothes and body parts in order to obtain a hierarchy of human abstraction. Alp Güler et al. (2018) collect a subset of Lin et al. (2014) in order to focus on person class by accurately setting pose, body parts among other tasks for multiple people. Another works also include motion/optical flow. Zhang et al. (2013) release a dataset for action recognition and tackle the problem training a collection of discriminative spatiotemporal patches based on temporal features and person joint coordinates, among others. Shahroudy et al. (2016) release a significant dataset surpassingly samples, human subjects, and camera views than previous ones and take into account 3D joints coordinates in order to enrich featuring for action classification by a part-aware LSTM framework. Furthermore, there are tasks involving 3D like body shape model, body parts shape segmentation, human 3D pose estimation. Ionescu et al. (2011) tackle human localization and 3D pose reconstruction by applying tractable augmented kernels to better encode complex dependencies among body parts and to reduce human pose search space. Ionescu et al. (2014) generate a large-scale 3D dataset by using motion capture and a set of body shape models to obtain multiple cues such as depth, 2D/3D joint coordinates, and body surface scan. They provide studies including nearest neighbor, standard linear/non-linear regression methods and kernels methods. Mehta et al. (2017) apply transfer learning from 2D to 3D human pose as a result of boosting performance on still images and generalize to a new dataset with a more extensive variety of real and augmented people views and appearances. Newell et al. (2016b) introduce a novel Convolutional Neural Networks (CNN) architecture to tackle 2D human pose that consists of intermediate supervision and skip connections in a stacked encoder-decoder fashion. As a consequence, they obtain very significant results and hence, that architecture is used in many works as the core network.

As it is common nowadays in most computer vision problems, deep learning, and particularly CNN, is the predominant methodology used by state of the art approaches. Outstanding results have been achieved by using deep learning in tasks like the 2D pose in the wild. However, other related tasks such as 3D pose, pixel-level segmentation, and human body depth estimation from RGB images still require further improvement in order to be accurately applied to real-world scenarios.

Recent approaches tend to benefit from unsupervised and cross-domain scenarios as in Zamir et al. (2018a) in order to reuse data and deal with related tasks by transfer learning. One standard technique in this scope is the use of multi-task approaches as in Everingham et al. (2015); Ionescu et al. (2011); Lin et al. (2014). Multi-task learning has been shown

to benefit human analysis tasks by leveraging the amount of data to be annotated since each image/video does not need a full annotation of all attributes: subsets of data can be annotated for different problems. Most importantly, while solving several tasks together, information is shared among them during training, providing them with complementary information for a better generalization.

In this chapter, we focus on multi-task learning of 2D pose, 3D pose, human body depth map, and body part segmentation from still images, which are common input cues for several human analysis tasks. We claim that these four tasks share semantic knowledge of the human body and, when jointly trained, can benefit each other for a better generalization. In particular, we extend the successful Hourglass network from Newell et al. (2016a) by learning each task as a separate stream and share information between tasks at different levels of the topology. Our contribution lies in the complementary analysis among the four main human body tasks on a multi-task setup. We evaluate which task combinations complement each other the best. To the best of our knowledge, this is the first time such a detailed analysis has been done in this domain.

To evaluate our framework, we focus on SURREAL from Varol et al. (2017), a synthetic dataset with real human bodies and annotations. Our results show that all four tasks benefit from the proposed multi-task module. We show some pairs of tasks do not help each other (e.g., 3D pose and body part segmentation), while others do so significantly (e.g., 2D pose and depth). Besides, multi-task learning provides higher performance improvements in those human body parts that show more variability in terms of spatial distribution, appearance and shape, e.g., wrists and ankles.

4.1.2 Related Work

The use of deep-learning techniques has been a breakthrough in most computer vision applications, including human analysis scenarios. Given the need for large volumes of data to train deep learning models, there is a recent trend in learning multi-task approaches. This paradigm shares information among different tasks for a better generalization, which can leverage the amount of annotated data required for each task.

Recent works like He et al. (2017); Kokkinos (2017); Omran et al. (2018); Varol et al. (2018) tend to extend the number of tasks to better benefit from sharing knowledge within cross-domain tasks. One extreme example can be found in He et al. (2017), where authors extend the number of tasks to eight, not just analyzing humans but objects and animals. Pyramid image decomposition is used as input to deal with semantic/boundary/object detection, normal estimation saliency/normal estimation, semantic/human part segmentation, semantic boundary detection, and region proposal generation. Other works like Dai et al. (2016); Luvizon et al. (2018); Popa et al. (2017); Zhao et al. (2018) add additional

tasks such as for instance segmentation, multi-human parsing, and mask segmentation. As an example, Dai et al. (2016) tackles instance segmentation, object detection and mask segmentation in a stacked fashion.

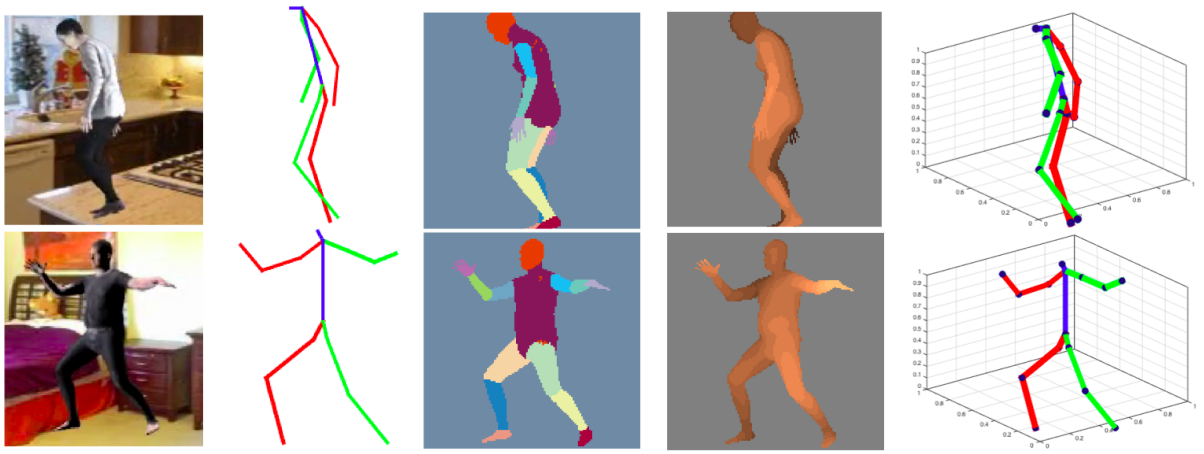
Different strategies exist in order to define multi-task schemes. Zamir et al. (2018a) perform a large-scale, cross-domain analysis on a new dataset of indoor scenes with no human interaction. They trained 26 neural networks, one per category and new combinations related to multiple domains via transfer learning. Most patterns found on this dataset exclude human kinematic constraints. Xia et al. (2017) build a two-stage FCN process that first detects human pose and then performs body parts parsing through a Conditional Random Field. The work of Alp Güler et al. (2018) uses Mask-RCNN from He et al. (2017) in a multi-task cascade fashion, connecting several intermediate layers for pose estimation and body parts parsing, while the Mask R-CNN from Kokkinos (2017) tackles instance/mask segmentation and object/key-point detection problems. Zhao et al. (2018) makes use of adversarial networks in a nested way, i.e., GAN outputs are used as the input to other GANs to deal with pose estimation and body parts parsing. In Popa et al. (2017); Wei et al. (2016) recursive processing stages are used to detect and segment 2d/3d pose and body-parts.

Another common combination of tasks is 2D/3D pose and body/clothes parsing from Ionescu et al. (2011) on datasets such as Pascal in Everingham et al. (2015) or COCO in Lin et al. (2014). The work of Nie et al. (2018) uses two encoders (2D pose and clothes parsing) with a module as a middle stream that acts as a parameter adapting to merge the features of both tasks and perform classification separately. In contrast, Liang et al. (2018) proposes a two-stage multi-task procedure that first uses a residual network to extract shared features. These are used by two CNN's performing 2D pose estimation and clothes parsing, respectively.

4.1.3 Multi-task human analysis

In this section, we first address the four selected tasks and then describe the proposed multi-task architecture for this analysis. We select four common tasks in many recent works: 2D/3D pose estimation, body parts segmentation, and body depth estimation. These tasks have some overlapping in the shared features/information, but each has a different definition: from depth or joints regression to pixel level classification. The goal is to design a compact model, consistent across tasks, such that overlapping features/information can be easily shared among all tasks in the model. By doing so, we can analyze which tasks are more correlated and in which parts we can achieve better improvement. The four tasks are described below.

- **2D pose**: This task tackles the estimation of 2D human joint coordinates. Heatmaps-



(a) RGB (b) 2D pose (c) Body Parts (d) Depth (e) 3D pose

Figure 4.1: Samples from SURREAL dataset with the chosen modalities.

based methods are state of the art for this task as in Newell et al. (2016a), consisting of estimating the location as Gaussian probability distribution around each joint. Each body joint is represented as a 2D heat map. These are stacked together, resulting in a 3D tensor where spatial relationships can be learned like Andriluka et al. (2014). In this method we use a tensor of size $64 \times 64 \times 16$, where $\#joints = 16$ (see Fig.4.1(b)).

- **Body parts segmentation**: The state-of-the-art on human body segmentation advocates training fully-convolutional networks that generate per pixel body part probabilities as in He et al. (2017); Xia et al. (2017). Body parts include hands, arms, legs, torso, and joints like ankles and knees. We define the segmentation output as a tensor of size $64 \times 64 \times 15$ where $\#parts = 14 + Background$ (see Fig.4.1(c)).

- **Full-body depth**: We tackle depth estimation as described in Haque et al. (2016), i.e. instead of regressing each pixel depth as a continuous value we quantize depth into $\#bins = 19$ bins resulting in a tensor of size $64 \times 64 \times (\#bins+1)$ (see Fig.4.1(d)). We define an extra bin for the background.

- **3D pose**: The standard approach for 3D pose estimation is coordinates regression as in Popa et al. (2017). However, regressing coordinates is highly non-linear and difficult to learn by a feature-coordinates mapping like Luvizon et al. (2018). Also, it is not consistent with other tasks. Following the heatmaps-based methods used in 2D pose estimation from Chen et al. (2017a), we use the target encoding used in Pavlakos et al. (2017); Varol et al. (2017, 2018). These works encode the 3D

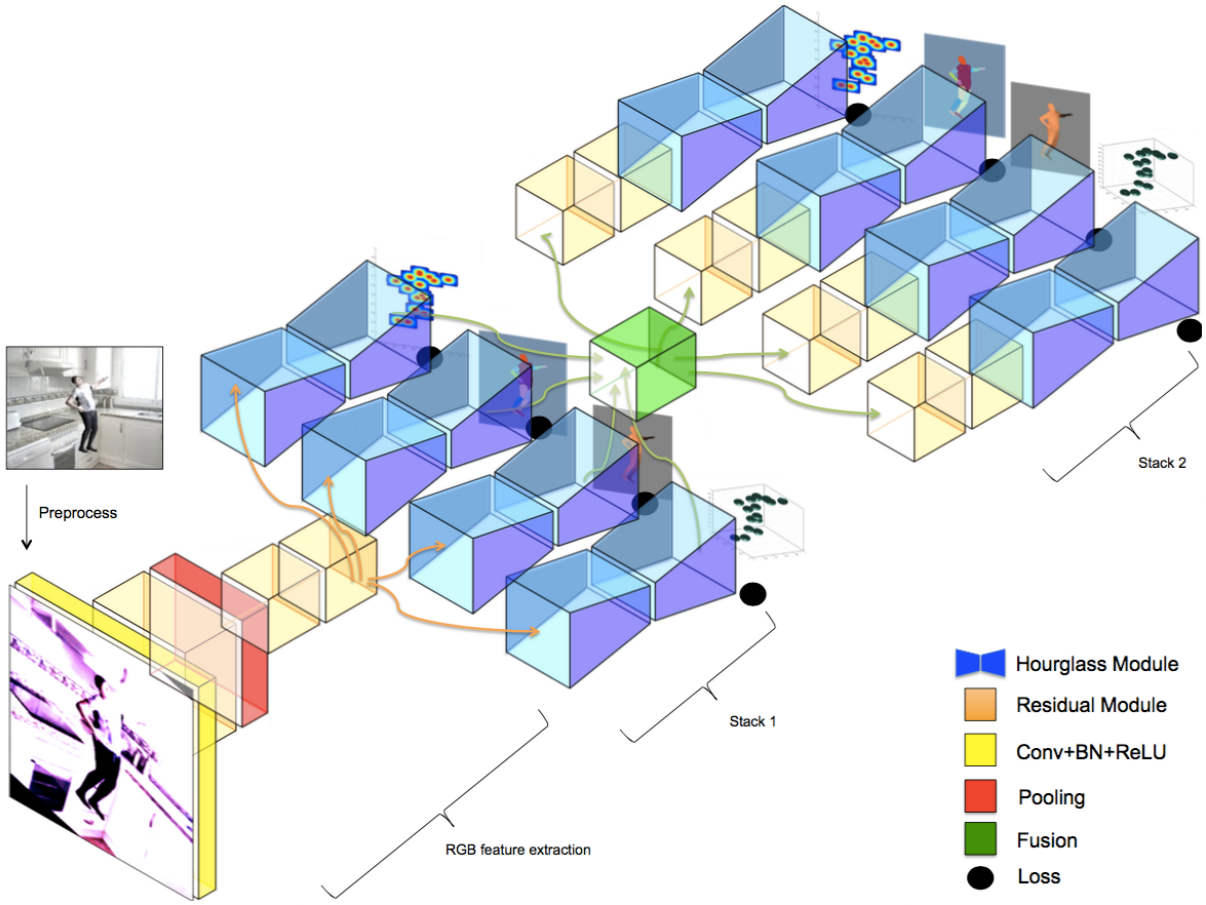


Figure 4.2: Proposed multi-task architecture.

location of the joints in the camera coordinate system like Luvizon et al. (2018) into 3D heat maps. 3D Gaussians are defined by a tensor of 3 dimensions for each joint (the same number of joints as in the 2D case) taking as referencing their corresponding 3D coordinates (see Fig.4.1(e)). The x and y axes are the standard Cartesian coordinates, being z-axis the depth as in the full-body depth estimation task. We output a tensor of size $64 \times 64 \times (\#bins \times \#joints)$ by binning depth information into 19 bins for each body part.

4.1.3.1 Multi-task architecture

We define all targets at the pixel level. Therefore any fully-convolutional deep architecture can be used for individual tasks. However, in this method we consider the Stacked Hourglass network (SH) from Newell et al. (2016a). This network has shown outstanding results for human pose estimation in still images. Each hourglass module consists of an encoder-decoder architecture with residual connections from encoder layers to corresponding decoder ones. The encoder consists of down-sampling residual modules that compress

the feature space in a latent representation tensor of size 4×4. The decoder contains up-sampling residual modules that enlarge the tensor to 64×64. The residual module includes several convolutional layers plus skips connections as in Kocabas et al. (2018). The skip connections from the encoder to decoder allows the model to fuse low-level features (e.g., edges, corners) with higher level features (e.g., semantics). The intermediate supervision at each hourglass module benefits from previous module outputs, refining and improving final network predictions. Given its high performance, its conceptual simplicity, and that allows for an easy multi-task integration among stacked modules, this architecture is serving as a baseline model in several works like Chen et al. (2017a); Ke et al. (2018); Luvizon et al. (2017); Ning et al. (2018); Yang et al. (2017).

In this method we use a stream, consisting of an SH network, to learn each task. These streams are then integrated by adding intermediate connectivity and supervision, as shown in Fig. 4.2. The resulting network is end-to-end trainable. Given an input RGB image, a set of residual modules are applied in order to generate shared features among all network streams (different tasks). Based on Newell et al. (2016a), several Hourglass modules can be stacked per stream. Each module has independent supervision and provides intermediate predictions as input to the next stacks. In our case, output features from each stream are concatenated to form a tensor of size $64 \times 64 \times (\#stream \times 256)$, where 256 is the default number of Hourglass features. Next, two residual modules are applied to each stream, the first convolving the joint features to the same feature space (standard practice as shown in Simonyan and Zisserman (2014)), and the second one compressing them to 256 features, again through convolution[1].

Regarding parameter estimation, a root-mean-square-error (RMSE) loss is used for 2D (L_{2Dpose}) and 3D (L_{3Dpose}) pose estimation, while cross-entropy (CE) across the spatial dimension of the heatmaps is used for depth estimation (L_{Depth}) and body part segmentation ($L_{BodyPart}$). Overall multi-task optimization is minimized by summing up the losses of all Hourglasses (4.1).

$$L_{Total} = L_{2Dpose} + L_{BodyPart} + L_{Depth} + L_{3Dpose} \tag{4.1}$$

4.1.4 Experiments

Here we describe the employed dataset, metrics and analysis of all four tasks, both standalone and multi-task networks.

[1]Note that our contribution in this chapter is not a design to compete with the state-of-the-art in each task, but rather a compact design to analyze cross-task contributions.

4.1.4.1 Data

In order to evaluate all multi-task combinations, we use SURREAL from Varol et al. (2017), a new large-scale dataset consisting of realistic synthetic data. The dataset is created by using recorded motion capture (MoCap) data to mimic realistic body movements in short video clips. The human body is rendered based on a body shape model. Then a cloth texture is added to the model including different lighting conditions. Finally, the model is projected to the image plane with a static background to have a realistic RGB image. The background is selected from indoor image datasets. Given this synthesis pipeline, different targets can be generated along with the RGB image: body depth maps, 2D/3D coordinates, body part segmentation, optical flow, and surface normals. The dataset contains nearly 6.5M frames. It consists of 145 subjects (115 train/30 test), 2,607 (1964 train/703 test) video sequences and 67,582 clips (55,001 train/12,528 test). Some samples are shown in Fig. 4.1 for the different data modalities.

4.1.4.2 Implementation details

We train different multi-task SH architectures considering different combinations of modalities to analyze their complementarity better. We train all models for 30 epochs using 2 Stacks of Hourglass, with a batch size of 5 and the RMSprop optimizer with learning rate $1e-3$. We first crop the image regions containing the centered human bodies using the provided bounding boxes of the dataset and resize them to 256×256 for training. Then, we apply standard data augmentation techniques such as scaling, jittering and rotation from Newell et al. (2016a). Moreover, the train/test splits are done such that 20% of the total is kept apart as in Varol et al. (2017).

In order to evaluate each modality, we make use of standard metrics: Intersection over Union (IOU) for body part segmentation, Percentage of Correct Keypoints thresholded at 50% of the head length (PCKh) as in Andriluka et al. (2014) for 2D pose estimation, root-mean-square-error (RMSE) for full body depth estimation and mean joint distance MJD in millimeters (mm) for 3D pose estimation. We also use success rate trend to analyze the evolution of the error/accuracy within different thresholds. This is given by the percentage of frames with an error smaller than the given thresholds.

4.1.4.3 Analysis of single-task models

Here we evaluate the models trained on specific tasks, which will serve as baselines to multi-task comparison.

4.1.4.3.1 Body part segmentation

The first column in Table 4.1 shows the single-task segmentation results, with an average IOU 67.48%. When looking at different body parts, the model shows high variability in accuracy: high performance for upper-body parts such as the head, torso, and legs, and lower performance for the feet, upper arms, and hands. This low accuracy in some parts (feet, hands) is due to these spanning just a few pixels, and regions of difficult interpretation, such as complex self-occlusions.

4.1.4.3.2 2D pose estimation

Regarding 2D pose estimation, the single-task 2D pose model already obtained an outstanding accuracy of 96.50% PCKh, as shown in Table 4.3. This may hint to the dataset being relatively simple for this kind of task, given the current state-of-the-art approaches. More specifically, we see lower accuracy on the wrists and elbows. These need a finer location since they have large scale variations. They may also be confused with the background on cluttered environments, depending on the clothing.

4.1.4.3.3 Full-body depth estimation

As shown in Table 4.4 the single-task depth model is capable of estimating the full-body depth (Mean Full Body row) with a 4.39% RMSE, a very low error. We can measure the depth prediction error on each body part by masking the predictions with the body part segmentation masks. Results obtained using only depth (Table 4.4, first column) show a higher error on hands and feet, and lower error on the torso, upper legs and upper arms due to their highly unconstrained kinematics in humans.

4.1.4.3.4 3D pose

In the case of 3D pose (Table 4.2, first column), we obtain an average error of 60.13mm, with the error being higher for the ankles and wrists. This is due to these covering a small spatial region, as well as corresponding to parts with many degrees of freedom. In summary, ankle, wrist, and elbow are the most difficult joints to learn. Again, we see those body parts and joints are difficult to predict for all tasks.

4.1.4.4 Analysis of multi-task models

The various considered tasks are highly related to each other and are based on similar visual cues. Thus, features extracted to solve a task may help to solve the others by providing a richer description of the body appearance. In this section, we evaluate how multi-task models help improve the accuracy of each individual task.

IOU	seg.	seg. + depth	3D pose + seg.	2D pose + seg.	2D pose + seg. + depth	2D/3D pose + seg.	3D pose + seg. + depth	2D/3D pose + seg. + depth
Background	98.0329	98.0726	98.0012	98.0631	**98.0781**	98.0641	98.0732	97.7579
Head	74.3689	74.4037	73.7297	74.2553	74.1704	74.3771	74.2328	**74.7454**
Torso	84.6390	84.8324	84.3057	**84.9853**	84.8013	84.9098	84.9153	80.6780
Upper R.Arm	65.8220	66.6616	65.8635	67.0540	66.1376	66.7216	66.4946	**67.7473**
Lower R.Arm	62.0338	62.5079	61.2258	62.6833	62.5857	62.1622	62.9103	**63.0192**
R. Hand	49.3243	48.4630	48.2553	49.4606	50.8114	48.5266	48.7750	**50.5932**
Upper L.Arm	65.4599	66.2077	65.8938	66.2191	65.3423	65.5865	65.8665	**66.7359**
Lower L.Arm	60.5462	61.4842	61.1205	61.1449	61.1934	60.5194	61.5981	**61.8868**
L.Hand	48.9188	48.9596	48.3028	46.8697	**49.2583**	46.1885	46.8591	48.6889
Upper R.Leg	75.2125	76.1054	75.1161	76.0184	**76.1120**	75.9127	75.8433	76.0172
Lower R.Leg	71.4514	72.2720	71.0750	71.9844	**72.3200**	71.7808	71.8441	71.9378
R.Feet	55.2237	55.5759	54.5336	55.4427	**56.7137**	54.4733	56.3635	55.8747
Upper L.Leg	75.2612	75.7805	75.4932	**76.2944**	76.2151	75.7273	76.1662	75.7628
Lower L.Leg	71.4049	72.2354	71.0951	72.2172	**72.3119**	71.5317	72.1089	71.5965
L.Feet	54.6201	55.1762	53.5035	53.9172	**56.4636**	53.6977	55.0263	54.9013
Mean	67.4880	67.9159	67.1677	67.7740	**68.1677**	67.3453	67.8051	67.8629

Table 4.1: Results on SURREAL dataset measuring body parts segmentation under IOU metric.

MJD (mm)	3D pose	2D/3D pose	3D pose + seg.	3D pose + depth	2D/3D pose + seg.	2D/3D pose + depth	3D pose + seg. + depth	2D/3D pose + seg. + depth
R.Ankle	86.1138	89.1075	81.6803	83.0312	87.9071	83.4697	**79.6775**	90.4500
R.Knee	59.9885	58.7382	**54.4890**	55.2095	56.9172	55.7307	55.1506	57.0098
R.Hip	25.6693	26.4384	25.7580	26.0962	26.5351	25.8593	25.5101	**25.4791**
L.Hip	25.4341	25.6198	25.7240	25.5058	26.2216	25.4403	25.5606	**25.0999**
L.Knee	56.9181	59.8854	56.5708	56.4425	58.4527	55.4873	**55.3666**	57.2093
L.Ankle	87.7192	89.7298	82.2631	84.6840	86.5020	83.1461	**81.0259**	87.6353
Thorax	31.2580	31.3804	31.4161	30.9884	31.1042	31.3439	30.5244	**30.0228**
Upper Neck	44.5032	42.5916	42.7647	42.3535	42.1803	42.7474	**41.2902**	42.2552
Head Top	49.6059	47.1529	46.9462	**46.7176**	49.1450	47.1224	46.8806	47.2783
R.Wrist	103.3092	103.4721	101.2466	107.9753	107.3964	105.2247	**100.6127**	102.6424
R.Elbow	70.3126	71.3751	70.3185	74.4315	72.5787	70.6057	**68.8732**	70.5880
R.Shoulder	46.1421	45.4316	46.1537	45.6363	45.7330	44.8304	**43.3576**	44.1882
L.Shoulder	47.4316	47.8410	45.2717	45.5204	46.2654	44.9013	**43.9592**	45.5271
L.Elbow	67.3347	68.6716	67.8447	68.5134	68.7963	67.6302	**63.9457**	65.2997
L.Wrist	100.3381	99.5600	96.5120	102.8758	103.0282	100.1062	**93.1507**	93.9053
Mean	60.1386	60.4664	58.3306	59.7321	60.5842	58.9097	**56.9924**	58.9727

Table 4.2: Results on SURREAL dataset measuring 3D pose under MJD (mm) metric.

4.1.4.4.1 Body Part Segmentation

As shown in Table 4.1, the tasks contributing the most to body part segmentation are 2D pose and depth estimation. Training a model to jointly solve these three tasks supposes a 1% improvement to the segmentation accuracy in terms of IOU (from 67.48% to 68.16%). Possible reasons are: 2D pose estimation may help to disambiguate pixel labels in the segmentation task by providing rough estimates of the body part locations; and depth estimation can help mitigating effects such as foreshortening, crowding and occlusion. Separately, both 2D pose and depth estimation improve the segmentation results relative to both IOU and pixel error.

Table 4.1 also shows that 3D body pose estimation is an inadequate complement for the segmentation task in terms of IOU. This may be due to the complexity of estimating the landmarks depth, with the model dedicating most of its capacity to this subtask. Moreover, the model encodes a relatively poor representation of the landmark locations in the image plane. This hypothesis is reinforced by the results of performing 2D+3D pose estimation along with body part segmentation. While 2D pose estimation does help the segmentation task, further adding 3D pose estimation results in worse accuracy than performing body part segmentation alone. The same effect happens with depth estimation and 3D pose. While depth estimation improves the overall segmentation accuracy, further performing 3D pose recovery results in worse accuracy.

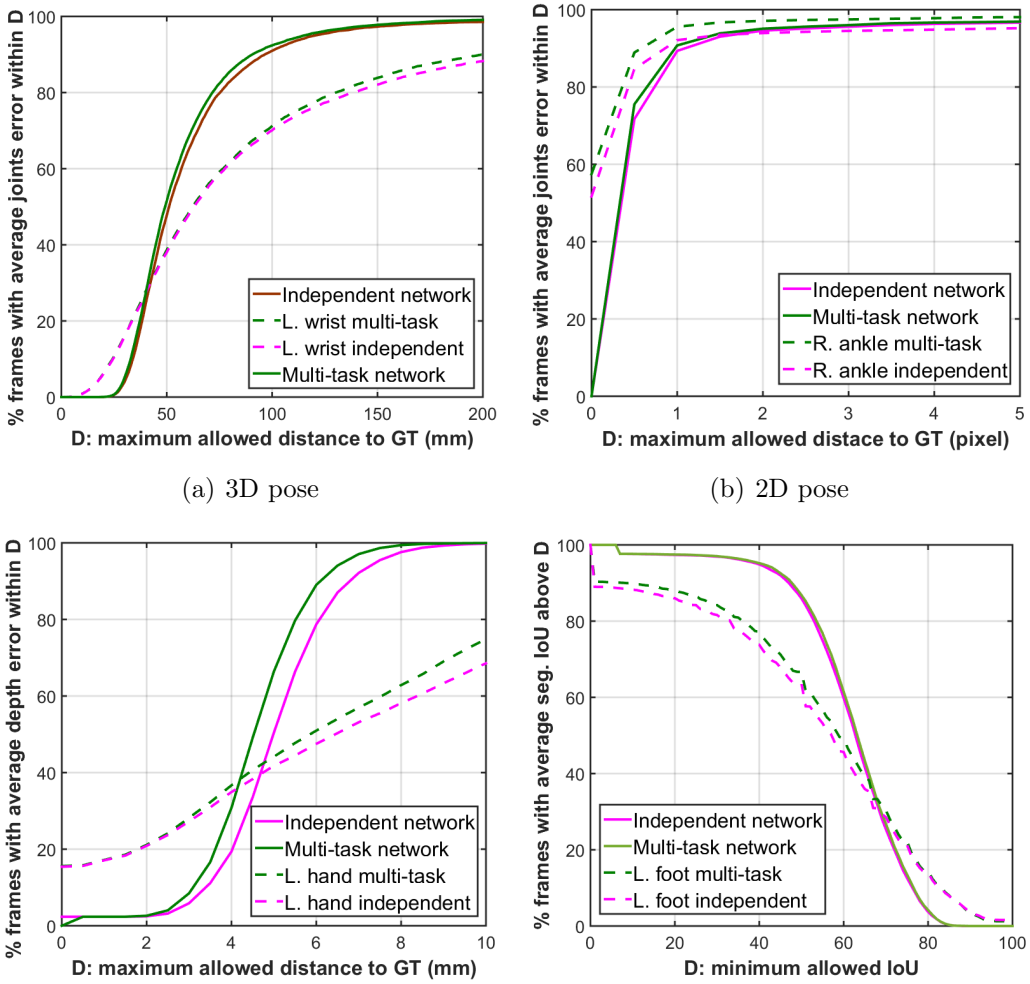
Looking at body parts results, one can see that performing 2D pose recovery along with body part segmentation improves IOU for the torso, arms, and legs. This is better reflected in the results for the model exploiting all considered subtasks. While adding 3D body pose recovery to the pipeline worsens the overall results of the best model, it does improve the segmentation accuracy of those parts it has been shown to improve on its own such as arms and hands.

Overall, we can say that the cues of 2D pose and depth estimation help to improve the segmentation accuracy. At the same time, 3D pose estimation worsens the overall results but helps improve the results for some specific body parts. The best overall model is found by performing 2D pose and depth estimation along with segmentation.

4.1.4.4.2 2D pose estimation

The results in Table 4.3 show the performance of the different multi-task models on 2D human pose estimation. We can see all task combinations improve on the single-task model, with the best results achieved by considering all tasks. Specifically, using all tasks results in a 0.51% improvement on the PCKh, going from 96.50% with the single-task model to 97.01% when using all tasks.

The single task contributing the most to 2D pose recovery is segmentation, resulting



(a) 3D pose

(b) 2D pose

(c) Full-body Depth map

(d) Body Parts segmentation

Figure 4.3: Success rate error for the different tasks. For each task: isolated task vs best multi-task approach; and for joint/part with highest multi-task improvement, its isolated task vs multi-task score.

PCKh	2D pose	2D pose + depth	2D/3D pose	2D pose + seg.	2D pose + seg. + depth	2D/3Dpose + seg.	2D/3D pose + depth	2D/3D pose + seg. + depth
R.Ankle	95.8064	95.8146	95.8064	96.3378	**96.8119**	96.6566	96.1007	96.6402
R.Knee	97.0326	96.959	96.91	97.2942	**97.4822**	97.1471	97.0817	97.4495
R.Hip	99.0109	99.1090	99.0272	99.1417	99.1090	99.0599	**99.1662**	99.0844
L.Hip	99.1253	99.2725	99.1008	99.2806	**99.3052**	99.2234	99.2970	99.2970
L.Knee	97.376	97.1552	97.2615	97.5721	**97.8256**	97.5149	97.5313	97.6457
L.Ankle	96.3214	96.0762	96.3132	96.8364	**97.0653**	96.8691	96.4686	96.9427
Pelvis	99.4687	99.5340	99.4932	99.5831	**99.6158**	99.4687	99.5586	**99.6158**
Thorax	99.3787	99.5095	99.4196	99.5014	99.5177	99.3951	99.5177	**99.5831**
Upper Neck	99.0763	99.1580	99.0763	99.1989	99.1171	99.0844	99.0763	**99.2234**
Head Top	98.7738	98.8556	98.8065	98.8229	98.8147	98.7329	98.8474	**98.9210**
R.Wrist	88.9970	89.0378	89.3567	89.9779	90.2068	89.8635	89.2586	**90.6237**
R.Elbow	94.3023	93.5339	94.0898	94.4004	**94.6211**	94.4167	93.9671	94.3268
R.Shoulder	98.3569	98.1362	98.3324	98.3978	98.5776	98.3733	98.2833	**98.7411**
L.Shoulder	98.0136	98.0544	97.9890	98.1852	98.2343	98.0953	97.8256	**98.4060**
L.Elbow	93.9099	94.2287	94.1061	94.6211	94.6129	94.6129	94.2042	**94.7029**
L.Wrist	89.1850	89.6264	89.3812	90.2232	90.5420	90.1659	90.2068	**91.0978**
Mean	96.5085	96.5039	96.5294	96.8360	96.9662	96.7925	96.6495	**97.0188**

Table 4.3: Results on SURREAL dataset measuring 2D pose under PCKh metric.

RMSE	depth	2D pose + depth	seg. + depth	3D pose + depth	2D pose + seg. + depth	2D/3D pose + depth	3D pose + seg. + depth	2D/3D pose + seg. + depth
Background	0.5151	**0.4955**	0.6372	0.5425	0.5727	0.6887	0.6830	0.5590
Head	4.7828	4.8319	4.5270	4.4978	4.5397	4.3778	**4.1174**	4.3523
Torso	2.7179	2.7216	2.4842	2.5810	2.538	2.5024	**2.3779**	2.5559
Upper R.Arm	3.8742	3.9463	**3.4306**	3.8128	3.5647	3.5505	3.4756	3.4641
Lower R.Arm	5.4385	5.4198	5.1384	5.3129	4.9385	5.1128	**4.8428**	5.0613
R. Hand	7.0447	7.0778	7.0683	6.9167	6.6483	6.8738	**6.6380**	6.9056
Upper L.Arm	3.7487	3.9582	3.4299	3.7295	3.5149	3.4873	**3.3240**	3.3965
Lower L.Arm	5.4778	5.6605	5.2851	5.4003	5.0954	5.1793	**4.8899**	5.1538
L.Hand	7.1597	7.2365	7.1001	6.9587	6.7485	6.9643	**6.6202**	6.9522
Upper R.Leg	3.3767	3.4739	3.2649	3.4919	3.2430	3.3522	**3.1933**	3.3732
Lower R.Leg	5.2455	5.3893	5.4107	5.3243	5.0982	5.1117	**4.8619**	5.1820
R.Feet	7.8622	7.9182	8.0064	7.9454	7.4462	7.7262	**7.3937**	7.7420
Upper L.Leg	3.3694	3.5158	3.2660	3.4426	3.2235	3.3606	**3.2014**	3.3314
Lower L.Leg	5.1918	5.4304	5.4314	5.3661	5.1769	5.1402	**4.9026**	5.1566
L.Feet	7.8774	8.0233	8.0535	7.9773	7.6125	7.8496	**7.5477**	7.8853
Mean Body Parts	4.9122	5.0066	4.8356	4.8867	4.6641	4.7518	**4.5378**	4.7381
Mean Full Body	4.3900	4.2300	4.3100	4.3500	4.1900	4.2500	**4.0400**	4.2400

Table 4.4: Results on SURREAL dataset measuring depth body parts estimation under RMSE metric.

in 0.3% increase. The said task may provide cues for the exact outline and localization of body parts, which can be easily leveraged for 2D body pose recovery. This is not the case of depth estimation, where body parts are not segmented. Still, depth estimation slightly improves the results, likely due to it providing an outline of the overall body, along with depth cues of the said outline, helping to disambiguate the location of the parts. 3D pose estimation, on the other hand, provides little complementary information about the location of the landmark relative to the camera plane, if any at all. If we look at individual joints, combining 2D pose, segmentation and depth improve on ankles and knees. Combining 2D/3D pose, segmentation and depth improves on the upper body and upper legs at the expense of losing precision on the other joints. This trade-off may be due to the ability of 3D pose estimation to disambiguate those joint locations suffering from cluttering and occlusions.

Summarizing, we see that performing all 4 tasks obtains the best results. By analyzing the other task combinations, we see that segmentation helps the most, followed by depth

estimation. Finally, 3D body pose estimation only helps marginally.

4.1.4.4.3 Full-body depth estimation

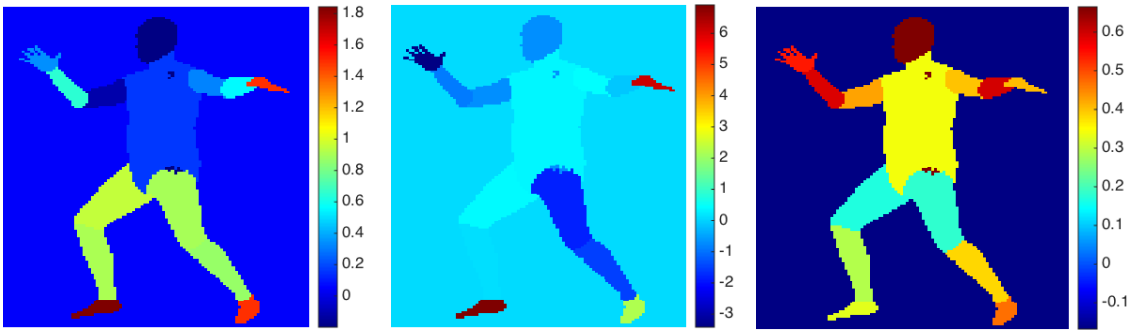
Here we evaluate the error on depth estimation for a collection of multi-task networks. Specifically, Table 4.4 shows that complementing depth estimation with 3D pose estimation and body part segmentation results in the best results: while the single-task model obtains a mean 4.39 RMSE calculated directly from the full-body depth prediction, the multi-task model goes down to an RMSE of 4.04, an 8% error reduction. Mean Body parts are the average of computing RMSE at each body part using its segmentation masks. Looking at tasks individually, segmentation contributes the most, with 3D pose estimation following closely. Segmentation may help depth estimation by providing richer semantic information on the body parts being segmented, allowing for a better model of the possible depth variability. On the other hand, 2D pose estimation does not contribute to solving the task, resulting in a higher error. This is due to this task not making use of depth information, resulting in bigger combined feature space with no additional depth cues in the encoding. We see this in higher order combinations: combining the successful tasks (segmentation and 3D pose estimation in addition to depth estimation) results in the best results. Further adding 2D pose estimation to the pipeline increases the overall error.

If we look at the results by a body part (Table 4.4), the best model, combining all tasks except for 2D pose estimation, obtains the lowest error in all cases. Compared to the baseline, some improvements to remark are the head, lower arms, and hands. This is due to the contribution of segmentation to better localize the parts layout and the 3D pose information to refine ambiguities at the depth level. Some difficult parts include the feet, lower legs, and hands.

4.1.4.4.4 3D pose estimation

This section analyzes the performance on 3D pose estimation of different multi-task models. Table 4.2 shows the prediction errors, in millimeters, for the different body joints and task combinations. The best overall results are obtained by considering the segmentation and depth estimation tasks along with 3D pose recovery, reducing the prediction error by 5% (from 60.13mm to 56.99mm).

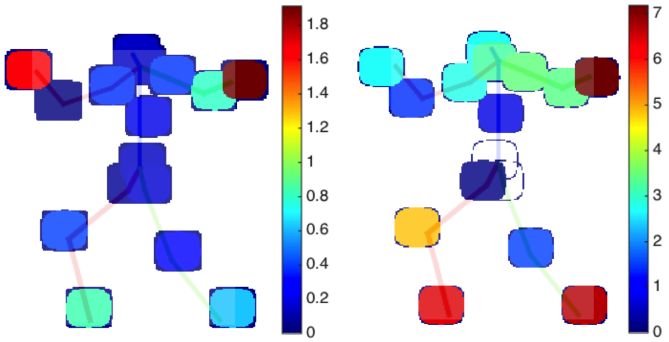
It is interesting to see that, similarly to 2D pose recovery, where 3D pose did not help improve the predictions, now it is the 3D pose that does not help. One can consider 2D pose recovery as a subtask of the 3D case, and thus the features used in 3D pose recovery already include those provided by the 2D case. In this case, the single task contributing the most to 3D pose recovery is segmentation, followed by depth estimation. This is



(a) Segmentation IOU: seg. vs 2D+seg.+depth

(b) Segmentation Pixel Accuracy: seg. vs 2D+seg+depth

(c) Full-body depth: depth vs 3D+seg.+depth



(d) 2D pose: 2D vs 2D/3D+seg.+depth

(e) 3D pose: 3D vs 3D+seg.+depth

Figure 4.4: Error visualization per each body part and task. The higher the value the higher the performance improvement for a particular metric of the best multi-task model compared to the baseline isolated task.

likely due to the same reasons discussed in the previous section: providing an outline of the body parts, and providing a general outline of the body with depth information, helping to disambiguate between parts during pose recovery.

Further combining both segmentation and depth estimation, as mentioned, obtains the best results, but not if we further consider 2D pose recovery. While in the previous section further adding 3D pose recovery to the 2D task did result in marginal benefits, in this case, there is no further information provided: 2D landmark localization is a problem already tackled when performing the same task in the 3D space. This results in slightly worse results when considering all tasks: a larger feature representation is provided but without encoding extra information, facilitating over-fitting.

If we inspect the results by the body joint, we find the best combination of tasks for most joints includes segmentation and depth to the 3D pose. On the other hand, hips and thorax also benefit from including 2D body pose information. This is likely due to

these parts forming the main portion of the body. A good 2D pose estimate may be more important for these parts since the ambiguity in depth is smaller. For parts with more depth uncertainty, like the ankles, knees, and wrists, considering 2D landmark estimation is highly detrimental to the 3D accuracy.

4.1.4.4.5 Analysis of success rate

We show success rate plots for different tasks in Fig. 4.3. For each modality, we compare independent SH network with the best multi-task network performing that task. We also show the trend for one of the parts that multi-task approach better improves, specifically left wrist for the 3D pose, Right ankle for the 2D pose, left a hand for depth map and left foot for part segmentation. As one can see in Fig. 4.3(c), full-body depth estimation benefits the most from multi-task learning, while 2D pose in Fig. 4.3(b) is the most accurate modality. In all cases, selected parts have higher than average gains for smaller error thresholds.

4.1.4.5 Comparison to the state-of-the-art

To the best of our knowledge, Varol et al. (2018) is the only state-of-the-art multi-task work evaluating on the SURREAL dataset. Similar to ours, they use SH modules to compute 2D/3D pose estimation and part segmentation. Differently, from us, 2D pose and body part segmentation are independent streams feeding information to the 3D pose stream. Full body depth estimation is not considered. We compare the results in Table 4.5. Note that we exclude background to compute segmentation IoU as in Varol et al. (2018). Unlike Varol et al. (2018) that trains 8 stacks for independent tasks and fine-tune 2 stacks in the multi-task model, we train our model from scratch using 2 stacks.

As one can see, our model is performing the best for 2D pose estimation in both independent and multi-task networks. Although our single-stream network performs better than Varol et al. (2018) in segmentation, our multi-task approach obtains similar results. In the case of 3D pose estimation, Varol et al. (2018) performs the best in both networks. Our multi-task network improves independent 3D pose by more than 3 mm while this improvement is 5.3 mm for Varol et al. (2018).

4.1.4.6 Discussion

This section summarizes some insights from the experiments performed for all tasks.

We have seen that at the 2D level cues from depth estimation are highly useful for both body parts segmentation and human pose recovery, while 3D pose estimation contributes marginally to the final performance. At the same time, body part segmentation and 2D pose estimation mutually benefit each other. Regarding body part segmentation, features

	Seg. (IoU)	2D pose (PCKh)	3D pose (MJD mm)
Varol *et al.* Varol et al. (2018) independent tasks	59.2	82.7	46.1
Varol *et al.* Varol et al. (2018) multi-tasks	**69.2**	90.8	40.8
Ours - independent tasks	65.3	96.5	**60.1**
Ours - multi-tasks	66.1	**97.0**	57.0

Table 4.5: State-of-the-art comparison on SURREAL.

from depth estimation improve the results the most, followed by 2D pose. Human pose recovery benefits from all other tasks, with the strongest cue being segmentation, followed by depth.

In contrast, at the 3D level, depth estimation and human pose recovery benefit from segmentation, similarly to the two 2D tasks. In contrast, 2D pose cues are the least relevant, since we can interpret the task as a subtask of 3D pose recovery. Both tasks use the same model to get the lowest error, that is, depth + segmentation + 3D pose. We argue this is due to segmentation enriching the representation with semantic cues, and the extra depth information either providing a more restrictive deformation model (3D pose estimation) or a more dense depth representation (body depth estimation).

Finally, a visual representation of the overall improvement of the best model per task and body part over the baseline are shown in Fig. 4.4. The higher the value the better average improvement for each particular task metric (e.g. 1.2 for a 3D joint represents an average improvement of 1.2 MJD error reduction). We can see in IOU and Pixel accuracy that parts with more degrees of freedom, such as feet, hands, and legs, are benefited the most from multi-tasking. In contrast, the trunk, head and upper arms, along with the background receive marginal improvements. For depth estimation, the improvements are more pronounced on the main body parts, such as the trunk and head, as well as the arms and hands. Then, for 2D and 3D pose, the former improved especially on the hands, while the latter improved on the upper body joints and ankles.

4.2 Conclusions

4.2.1 Multi-task human analysis in still images: 2D/3D pose, depth map, and multi-part segmentation

We analyzed the multi-tasking paradigm on four human body problems: 2D/3D body pose estimation, full-body depth estimation, and body parts segmentation. We concluded that each task benefits each other at some ratios an aspects. Depth estimation and body part segmentation help each other, while 2D/3D pose estimation benefit mainly from the segmentation one. Depth helps to disambiguate body parts, while segmentation provides more robust region context for joints localization. However, very related tasks such as 3D pose and 2D pose do not take benefit of each other since the latter can be contextualized as a subtask of the 3D pose.

Chapter 5

Conclusions and future research

5.1 Conclusions

5.1.1 HuPBA 8k+: Dataset and ECOC-GraphCut based Segmentation of Human Limbs

In this chapter, we introduced the *HuPBA* $8K+$ dataset, which to the best of our knowledge is the most significant multi-limb RGB dataset for Pose Recovery, with more than 120 000 manually labeled limb regions. Besides, we proposed a novel two-stage method for human multi-limb segmentation in RGB images. In the first stage, we perform a person/background segmentation by training a set of body parts using cascades of classifiers embedded in an ECOC framework. In the second stage, to obtain a multi-limb segmentation we applied multi-label GraphCuts to a set of limb-like probability maps obtained from a more powerful problem-dependent ECOC scheme.

We compared our proposal with state-of-the-art pose-recovery approaches on the novel dataset obtaining very satisfying results in terms of both person/background and multi-limb segmentation. For completeness, the novel dataset was also labeled with different human actions drawn from a 11 gesture dictionary. In this sense, we also provide with gesture recognition results as a firm baseline to share with the Computer Vision community.

5.1.2 Learning to segment humans by stacking their body parts

We presented a two-stage scheme based on the MSSL framework for the segmentation of the human body in still images. We defined an extended feature set by stacking a multi-scale decomposition of body part likelihood maps, which are learned employing a multi-class classifier based on soft body part detectors. The extended set of features

encodes spatial and contextual information of human limbs which combined enabled us to define features with high order information. We tested our proposal on a large dataset obtaining significant segmentation improvement over state-of-the-art methodologies. As future work, we plan to extend the MSSL framework to the multi-limb case, in which two multi-class classifiers will be concatenated to obtain a multi-limb segmentation of the human body that takes into account contextual information of human parts.

5.1.3 Multi-task human analysis in still images: 2D/3D pose, depth map, and multi-part segmentation

In this chapter, we analyzed the contribution of multi-tasking on four standard bodies pose analysis problems: 2D/3D body pose recovery, full-body depth estimation, and body parts segmentation. We have found that problems looking at complementary aspects of the problem benefit each other the most. Depth estimation and body part segmentation help each other, while 2D/3D body poses estimation benefit mainly from body part segmentation, followed by depth estimation. These tasks provide complementary features: depth information helps disambiguate body parts, while body part segmentation provides more robust features for locating joints during body pose estimation. Also, 3D pose estimation helps depth estimation, likely by reducing ambiguity: 3D pose estimation helps to restrict the space of possible body poses. On the other hand, features from problems that are too closely related do not help significantly improve the predictions: 3D pose recovery already includes the 2D problem as a subtask, already encoding its features. For 2D pose recovery, features coming from the 3D case sacrifice precision in the camera plane, allotting more network capacity to estimate the landmarks depth.

5.2 Possible directions for future research

Different possibilities for future research in the different contributions of this thesis are discussed in the following sections.

5.2.1 HuPBA 8k+: Dataset and ECOC-GraphCut based Segmentation of Human Limbs

As a future perspective, it would be interesting to exploit the dataset key features, in order for the Computer Vision community to use in every possible manner (comparisons, validation, challenges, etc.). Furthermore, the experimental results obtained encourages us to follow this line by making use of more advanced techniques for both multi-class classification (taking into account contextual information and relative spatial relations)

and multi-label segmentation. The proposed ensemble strategy is independent of the base classifier considered for body-part estimation. In this sense, future work includes the adaptation of more accurate deep learning approaches and multi-task ones, as the one presented in the last chapter, within the ensemble strategy proposed in this work for human body segmentation.

5.2.2 Learning to segment humans by stacking their body parts

As future work, one way is to extend the MSSL framework to the multi-limb case, in which two multi-class classifiers will be concatenated to obtain a multi-limb segmentation of the human body that takes into account contextual information of human parts. In the same way as previous work, stacked learning is also independent of the primary classifier to be stacked. One example is the Hourglass model used in the last chapter. Thus, future work also includes the adaptation of more accurate deep detectors to be stacked using the methodology used in this work.

5.2.3 Multi-task human analysis in still images: 2D/3D pose, depth map, and multi-part segmentation

Several research lines are open to work in the future. First, based on the research of Zamir et al. (2018b), we can study how the pre-trained models learned from the 26 categories of indoor scenarios can contribute if fused with our multi-task model. We must take into account that most patterns found on taskonomy dataset exclude human kinematic constraints. Thus, we could use both approaches to model human-scene interaction. Second, we face training on synthetic data, and some research has already been done to study the gap between synthetic and real data. However, few works have carried out human body analysis using pre-trained models on synthetic data to be applied to real data. In this sense, the study of common latent spaces of real and synthetic data to benefit human pose estimation and segmentation can be considered as future work. Third, it is interesting to study a different kind of architectures to perform a more detail analysis and an interpretation of network topology/structure concerning generalization capability to human pose and segmentation problems.

Bibliography

A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):44–58, 2006.

R. Alp Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Computer Society Conference on*, pages 1014–1021. IEEE, 2009.

M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 623–630. IEEE, 2010.

M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

S. Baluja and R. Caruana. Removing the genetics from the standard genetic algorithm. In A. Prieditis and S. Russel, editors, *The Int. Conf. on Machine Learning 1995*, pages 38–46, San Mateo, CA, 1995. Morgan Kaufmann Publishers.

M. Á. Bautista, S. Escalera, X. Baró, P. Radeva, J. Vitriá, and O. Pujol. Minimal design of error-correcting output codes. *Pattern Recognition Letters*, 33(6):693–702, 2012a.

M. A. Bautista, S. Escalera, X. Baró, P. Radeva, J. Vitriá, and O. Pujol. Minimal design of error-correcting output codes. *Pattern Recogn. Lett.*, 33(6):693–702, Apr. 2012b. ISSN 0167-8655.

M. Á. Bautista, S. Escalera, X. Baró, and O. Pujol. On the design of an ecoc-compliant genetic algorithm. *Pattern Recognition*, 47(2):865–884, 2014.

M. A. Bautista, A. Hernández-Vela, S. Escalera, L. Igual, O. Pujol, J. Moya, V. Violant, and M. T. Anguera. A gesture recognition system for detecting behavioral patterns of adhd. *IEEE transactions on cybernetics*, 46(1):136–147, 2015.

H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

C. Blakemore and G. F. Cooper. Development of the brain depends on the visual environment. *Nature*, 228(5270):477, 1970.

L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Computer Vision (ICCV), 2019 IEEE International Conference on*, pages 1365–1372. IEEE, 2009.

L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *European conference on computer vision*, pages 168–181. Springer, 2010.

Y. Boykov and G. Funka-Lea. Graph cuts and efficient nd image segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006.

Y. Boykov and V. Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. In *Computer Vision and Pattern Recognition, 2003. CVPR 2003. IEEE Computer Society Conference on*, pages 26–33, 2003.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.

L. Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

J. Canny. A computational approach to edge detection. In *Readings in computer vision*, pages 184–203. Elsevier, 1987.

V. Carvalho and W. Cohen. Stacked sequential learning. *Proceedings of the IJCAI-05, Edinburgh, Scotland*, 2005.

B. Chakraborty, A. D. Bagdanov, J. Gonzalez, and X. Roca. Human action recognition using an ensemble of body-part detectors. *Expert Systems*, 30(2):101–114, 2013.

Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1212–1221, 2017a.

Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *CoRR, abs/1705.00389*, 2, 2017b.

Y.-T. Chen and C.-S. Chen. Fast human detection using a novel boosted cascading structure with meta stages. *Image Processing, IEEE Transactions on*, 17(8):1452–1464, 2008.

C.-J. Chou, J.-T. Chien, and H.-T. Chen. Self adversarial training for human pose estimation. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 17–30. IEEE, 2018.

X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2017.

D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*, 2011.

W. W. Cohen. Stacked sequential learning. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2005.

J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005a.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886 –893 vol. 1, 2005b.

M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3041–3048, 2013.

F. De la Torre, J. K. Hodgins, J. Montano, and S. Valcarcel. Detailed human data acquisition of kitchen activities: the cmu-multimodal activity database (cmu-mmac). Technical report, RI-TR-08-22h, CMU, 2008.

T. Dietterich. Machine learning for sequential data. *Lecture Notes in*, 2002.

T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence research*, 2:263–286, 1994.

K. Dinakar, E. Weinstein, H. Lieberman, and R. L. Selman. Stacked generalization learning to analyze teenage distress. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *European conference on computer vision*, pages 228–242. Springer, 2010.

M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International journal of computer vision*, 99(2):190–214, 2012.

M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12): 2179–2195, 2009.

S. Escalera, D. M. Tax, O. Pujol, P. Radeva, and R. P. Duin. Subclass problem-dependent design for error-correcting output codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):1041–1054, 2008.

S. Escalera, O. Pujol, and P. Radeva. On the decoding process in ternary error-correcting output codes. *PAMI*, 32:120–134, 2010a.

S. Escalera, O. Pujol, and P. Radeva. On the decoding process in ternary error-correcting output codes. *IEEE transactions on pattern analysis and machine intelligence*, 32(1): 120–134, 2010b.

S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. *ChaLearn Multi-modal Gesture Recognition Grand Challenge and Workshop, 15th ACM International Conference on Multimodal Interaction*, 2013.

S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014:

Dataset and results. In *European Conference on Computer Vision*, pages 459–473. Springer, 2014.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, June 2010.

M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 66–73. IEEE, 2000.

P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

P. F. Felzenszwalb and D. McAllester. Object detection grammars. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, volume 18, 2011.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, 100(1):67–92, 1973.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT*, pages 23–37, 1995.

C. Gatta, E. Puertas, and O. Pujol. Multi-scale stacked sequential learning. *Pattern Recognition*, 44(10-11):2414–2426, 2011.

D. M. Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

R. B. Girshick, P. F. Felzenszwalb, and D. A. Mcallester. Object detection with grammar models. In *Advances in Neural Information Processing Systems*, pages 442–450, 2011.

G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik. Articulated pose estimation using discriminative armlet classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3342–3349, 2013.

K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Computer Vision and Pattern Recognition, 2017. CVPR 2017. IEEE Computer Society Conference on*, pages 6757–6765. IEEE, 2017.

A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In *European Conference on Computer Vision*, pages 160–177. Springer, 2016.

C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988.

K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

A. Hernández-Vela, M. Reyes, V. Ponce, and S. Escalera. Grabcut-based human segmentation in video sequences. *Sensors*, 12(11):15376–15393, 2012a.

A. Hernández-Vela, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov, and S. Escalera. Graph cuts optimization for multi-limb human segmentation in depth maps. In *Computer Vision and Pattern Recognition, 2012. CVPR 2012. IEEE Computer Society Conference on*, pages 726–732, 2012b.

A. Hernández-Vela, M. Á. Bautista, X. Perez-Sala, V. Ponce-López, S. Escalera, X. Baró, O. Pujol, and C. Angulo. Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in rgb-d. *Pattern Recognition Letters*, 2013.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591, 1959.

D. H. Hubel and T. N. Wiesel. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *The Journal of physiology*, 206(2):419–436, 1970.

C. Ionescu, F. Li, and C. Sminchisescu. Latent structured models for human pose estimation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2220–2227. IEEE, 2011.

C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2014.

S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1233–1258, 1987.

M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

L. Ke, M.-C. Chang, H. Qi, and S. Lyu. Multi-scale structure-aware network for human pose estimation. In *European Conference on Computer Vision*, pages 713–728, 2018.

M. Kocabas, S. Karagoz, and E. Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *European Conference on Computer Vision*, pages 417–433, 2018.

I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6129–6138, 2017.

E. B. Kong and T. G. Dietterich. Error-correcting output coding corrects bias and variance. In *Machine Learning Proceedings 1995*, pages 313–321. Elsevier, 1995.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

L. Ladicky, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3578–3585, 2013.

I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Computer Vision and Pattern Recognition, 2017. CVPR 2017. IEEE Computer Society Conference on*, pages 4704–4713. IEEE, 2017.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

M. K. Leung and Y.-H. Yang. Human body motion segmentation in a complex scene. *Pattern recognition*, 20(1):55–64, 1987.

M. K. Leung and Y.-H. Yang. First sight: A human body outline labeling system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(4):359–377, 1995.

W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

X. Liang, K. Gong, X. Shen, and L. Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional network. In *Proceedings of the IEEE international conference on computer vision*, pages 2648–2655, 2013.

Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang. Macro-micro adversarial network for human parsing. In *European Conference on Computer Vision*, pages 418–434, 2018.

D. C. Luvizon, H. Tabia, and D. Picard. Human pose regression by combining indirect part detection and contextual information. *arXiv preprint arXiv:1710.02322*, 2017.

D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Computer Vision and Pattern Recognition, 2018. CVPR 2018. IEEE Computer Society Conference on*, volume 2, 2018.

D. Marr and E. Hildreth. Theory of edge detection. *Proc. R. Soc. Lond. B*, 207(1167): 187–217, 1980.

D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 International Conference on*, pages 506–516. IEEE, 2017.

K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, pages 69–82. Springer, 2004.

G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2004.

D. Munoz, J. A. Bagnell, and M. Hebert. Stacked hierarchical labeling. In *European Conference on Computer Vision*, pages 57–70. Springer, 2010.

A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016a.

A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016b.

A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016c.

X. Nie, J. Feng, J. Xing, and S. Yan. Generative partition networks for multi-person pose estimation. *arXiv preprint arXiv:1705.07422*, 2017.

X. Nie, J. Feng, Y. Zuo, and S. Yan. Human pose estimation with parsing induced learner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2100–2108, 2018.

G. Ning, Z. Zhang, and Z. He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, 20(5):1246–1259, 2018.

G. L. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox. Deep learning for human part discovery in images. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1634–1641. IEEE, 2016.

M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018.

G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Computer Vision and Pattern Recognition, 2017. CVPR 2017. IEEE Computer Society Conference on*, pages 1263–1272. IEEE, 2017.

L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013a.

L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *Proceedings of the IEEE international conference on Computer Vision*, pages 3487–3494, 2013b.

A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *Conference on Computer Vision and Pattern Recognition*, volume 1, page 5, 2017.

E. Puertas, M. Bautista, D. Sanchez, S. Escalera, and O. Pujol. Learning to segment humans by stacking their body parts. In *European Conference on Computer Vision*, pages 685–697. Springer, 2014.

E. Puertas, S. Escalera, and O. Pujol. Generalized multi-scale stacked sequential learning for multi-class classification. *Pattern Analysis and Applications*, 18(2):247–261, 2015.

V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *European Conference on Computer Vision*, pages 33–47. Springer, 2014.

D. Ramanan. Learning to parse images of articulated bodies. In *Advances in neural information processing systems*, pages 1129–1136, 2006.

D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 271–278. IEEE, 2005.

D. Ramanan, D. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *PAMI*, 29(1):65 –81, jan. 2007.

M. Reyes, G. Dominguez, and S. Escalera. Featureweighting in dynamic timewarping for gesture recognition in depth data. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1182–1188. IEEE, 2011.

C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, Aug. 2004a. ISSN 0730-0301.

C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004b.

B. Rothrock, S. Park, and S.-C. Zhu. Integrating grammar and segmentation for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3214–3221, 2013.

H. Sakoe, S. Chiba, A. Waibel, and K. Lee. Dynamic programming algorithm optimization for spoken word recognition. *Readings in speech recognition*, 159:224, 1990.

D. Sánchez, J. C. Ortega, M. Á. Bautista, and S. Escalera. Human body segmentation with multi-limb error-correcting output codes detection and graph cuts optimization. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 50–58. Springer, 2013.

D. Sánchez, M. Á. Bautista, and S. Escalera. Hupba8k+: Dataset and ecoc-graph-cut based segmentation of human limbs. *Neurocomputing*, 150:173–188, 2015.

B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *Computer Vision and Pattern Recognition, 2010. CVPR 2010. IEEE Computer Society Conference on*, pages 422–429. IEEE, 2010a.

B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *European conference on computer vision*, pages 406–420. Springer, 2010b.

B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *Computer Vision and Pattern Recognition, 2011. CVPR 2011. IEEE Computer Society Conference on*, pages 1281–1288. IEEE, 2011.

C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion-Based Recognition*, pages 227–243. Springer, 1997.

W. Sun. A stacked sub-word model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1385–1394. Association for Computational Linguistics, 2011.

W. Tang, P. Yu, and Y. Wu. Deeply learned compositional models for human pose estimation. In *European Conference on Computer Vision*, pages 190–206, 2018.

K. M. Ting and I. H. Witten. Stacked generalization: when does it work? 1997.

K. M. Ting and I. H. Witten. Issues in stacked generalization. *Journal of artificial intelligence research*, 10:271–289, 1999.

D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *European Conference on Computer Vision*, pages 227–240. Springer, 2010.

G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.

G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. *arXiv preprint arXiv:1804.04875*, 2018.

V. Vineet, J. Warrell, L. Ladicky, and P. Torr. Human instance segmentation from video using detector-based conditional random fields. In *BMVC*, 2011.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. IEEE Computer Society Conference on*, volume 1, pages 511–518, 2001a.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001b.

C. Wang and S. Mahadevan. Manifold Alignment Preserving Global Geometry. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI'13, pages 1743–1749. AAAI Press, 2013.

Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *Computer Vision and Pattern Recognition, 2011. CVPR 2011. IEEE Computer Society Conference on*, pages 1705–1712. IEEE, 2011.

S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding*, 115(2):224–241, 2011.

Wikipedia. Herakles and athena, 2007. URL `https://upload.wikimedia.org/wikipedia/commons/4/4f/Athena_Herakles_Staatliche_Antikensammlungen_2648.jpg`. [Online; accessed September 30, 2007].

Wikipedia. Cave paintings, 2011. URL `https://en.wikipedia.org/wiki/Cave_painting#/media/File:Altamira_bisons.jpg`. [Online; accessed April 8, 2011].

Wikipedia. Egyptian farmers, 2016a. URL `https://upload.wikimedia.org/wikipedia/commons/2/27/Egyptian_Farmers.jpg`. [Online; accessed June 16, 2016].

Wikipedia. War panel, 2016b. URL `https://upload.wikimedia.org/wikipedia/commons/f/f9/Standard_of_Ur_-_War.jpg`. [Online; accessed February 27, 2016].

D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

F. Xia, P. Wang, X. Chen, and A. L. Yuille. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6769–6778, 2017.

B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision*, pages 466–481, 2018.

J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Computer Society Conference on*, pages 1794–1801. IEEE, 2009.

W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1281–1290, 2017.

Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1385–1392. IEEE, 2011.

Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2013.

B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9–16. IEEE, 2010.

C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th annual international conference on machine learning*, pages 1169–1176. ACM, 2009.

A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018a.

A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018b.

H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, and J. Jia. Human pose estimation with spatial contextual information. *arXiv preprint arXiv:1901.01760*, 2019.

W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2248–2255, 2013.

J. Zhao, J. Li, Y. Cheng, L. Zhou, T. Sim, S. Yan, and J. Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. *arXiv preprint arXiv:1804.03287*, 2018.

F. Zhou, F. De la Torre, and J. K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):582–596, 2013.

Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1491–1498. IEEE, 2006.