SAPIENZA
UNIVERSITÀ DI ROMA

Towards An Understanding Of Human Activities: From The Skeleton To The Space

Department of Computer, Control, and Management Engineering

Ph. D. in Engineering in Computer Science – XXXI Cycle

Candidate

Marta Sanzari
ID number 1211716

Thesis Advisor

Prof. Fiora Pirri

Co-Advisor

Prof. Marco Schaerf

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Engineering in Computer Science

April 2019

Abstract

In this thesis is described the reasearch undertaken for the Ph.D. project in Computer Vision, having the main objective to tackle human activity recognition from RGB videos.

Human activity recognition from videos aims to recognize which human activities are taking place during a video, considering only cues directly extracted from video frames. The related applications are manifold: healthcare monitoring applications, such as rehabilitation or stress monitoring, monitoring and surveillance for indoor and outdoor activities, human-machine interaction, entertainment etc..

An important disambiguation has to be exposed before proceeding further: the one between action and activity. Actions are generally described in literature as single person movements that may be composed of multiple simple gestures organized temporally, such as walking, waving or and punching. Gestures are instead elementary movements of a body part. On the other hand, activities are described as involving two or more persons and/or objects, or a single person performing complex actions, i.e. a sequence of actions.

Human activity recognition is one of the main subjects of study of computer vision and machine learning communities since a long time, and it is still an hot topic due to its complexity.

A challenging task is to develop a system for human activity recognition, due to well-known computer vision problems. Body parts occlusions, light conditions, and image resolution are only a subset of this problems. Furthermore, similitudes between activity classes make the problem even harder. Activities in the same class may be exhibited by distinct persons with distinct human body movements, and activities in different classes may be hard to discriminate because they may be constituted by analogous information. The way in which humans execute an activity depends on their habits, and this drives the challenge of detecting activities quite difficult.

The main consideration coming out deeply analyzing the available literature for activity recognition, is that an activity recognition robust system has to be context-aware. Namely, not only the human motion is important to achieve good performances, but also other relevant cues which can be extracted from videos have to be considered.

The available state of the art research in computer vision still misses a complete framework for human activity recognition based on context, taking into account both the scene where activities are taking place, objects analysis, 3D human motion analysis and interdependence between activity classes. This thesis describes computer vision frameworks which will enable the robust recognition of human activities explicitly considering the scene context.

In this thesis are described the main contributions for context-aware activity recognition regarding 3D modeling of articulated and complex objects, 3D human pose estimation from single images and a method for activity recognition

based on human motion primitives. Four major publications will be presented, together with an extensive literature review concerning computer vision areas such as 3D object modeling, 3D human pose estimation, human action recognition, human action recognition based on action and motion primitives and human activity recognition based on context. Future work concerning the undertaken research will be to build a complete system for activity recognition based on context, exploiting the several frameworks introduced so far.

Keywords: Computer Vision, Human Activity Recognition, 3D object Modeling, 3D Human Pose Estimation.

Acknowledgments

Completion of this Doctoral research would not have been possible without the support and assistance of numerous people throughout the research project. I would like to express my appreciation to my Principal Supervisor, Prof. Fiora Pirri, and to my colleagues. Their support, guidance and professional advice provided to me throughout the duration of the research has been invaluable and I am extremely grateful for their assistance. I would like to express my gratitude to my parents, who always supported me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The aim of the Ph.D. project is to create a framework for the recognition of human activities from videos, based on human motion and on other relevant cues present in the scene such as objects or the environment in which the activity is taking place.

If for example we want to recognize the activity 'Springboard diving' from an RGB video, beyond the motion of the person who is diving we can also find in the scene relevant features for the recognition of this specific activity such as the presence of the springboard, of the swimming pool or the bleachers. Another relevant cue could be the interaction between the person and the springboard.

The main consideration coming out deeply analyzing the available literature for activity recognition, is that an activity recognition robust system has to be context-aware. Namely, not only the human motion is important to achieve good performances, but also other relevant cues which can be extracted from videos have to be considered.

The available state of the art research in computer vision still misses a complete framework for human activity recognition based on context, taking into account both the scene where activities are taking place, objects analysis, 3D human motion analysis and interdependence between activity classes. This thesis describes computer vision frmeworks which will enable the robust recognition of human activities explicitly considering the scene context.

In order to succeed in this project we are working on different computer vision problems. The initial research covered 3D object modeling from few or single images, both of simple objects such as convex ones (Ntouskos et al., 2015b) and of complex objects such as concave or with reflective surfaces (Natola et al., 2016). 3D modeling of articulated objects was obtained modeling the single aspects of object components (i.e. object main parts), and then assembling them together. 3D modeling of complex objects was obtained collecting a

database of 3D objects and using normal field information to learn a dictionary describing the correspondence between visual features and 3D normal field. The normal field is in turn used to model the 3D shape of new unseen objects. The motivation of these researches related to the goal of the project is that the 3D shape of objects involved in the scene can highlight relevant information for the recognition of object manipulability, this is the reason for our studies on objects affordance (Sanzari et al., 2015), and especially can help to understand which human activities are performed allowing to capture the interaction between persons and objects.

The second step of the Ph.D. research project was dealing with action recognition from 3D MoCap (Motion Capture) data, in particular we focused on recognizing specific patterns relative to the 3D motion of different human body parts such as the arms, the legs, the torso and the head. By representing configurations of actions as manifolds, joint positions are mapped on a subspace via principal geodesic analysis. The reduced space is highly informative and allows for classification based on a non-parametric Bayesian approach, generating behaviors for each sub-body part. Classifying these patterns specific for different kind of actions we were able to obtain action recognition (Natola et al., 2015b).

Facing the problem of action recognition based on 3D human skeleton joints data, it was clear that it is not possible to deal only with 3D data for our approaches. For this reason we focused on the problem of 3D human pose estimation from single images (Sanzari et al., 2016). Through a good estimation of 3D poses of human skeleton joints we are able to use not only databases of 3D human motion, or special and expensive equipment such as Vicon, to face the problem of action recognition. We are able now to obtain 3D data directly from RGB images and videos.

Currently we are working on two different problems: the discovery of human motion primitives for the recognition and classification of human motion, and the recognition of human activities based on context.

We are facing the problem of recognizing human motion building a robust framework able to deal with the well-known problems related to human motion, such as body parts occlusions. For this purpose we are analyzing the discovery and recognition of human motion primitives from 3D skeleton data (Sanzari et al., 2019), which are those movements that span an interval of time in which a change in position of a limb or body part takes place. Human motion primitives are discovered by optimizing the 'motion flux', a quantity which depends on the motion of a group of human skeleton joints. Motion primitives are recognized analyzing the geometric features belonging to different primitive movements. The discovered motion primitives are in turn used to identify human activities.

For the recognition of human activities based on context, the future work will concern the development of a framework for activity recognition from RGB videos based on the scene, the human motion and the presence in the scene of

relevant objects, for example manipulated ones.

1.2 Thesis Structure

The research described in this thesis details the processes adopted for tackling different computer vision problems needed to build the ground for a framework to understand human activities from RGB videos based on context.

This involved detailed investigations into the resolution of some of the inherent deficiencies discovered in the available approaches in the computer vision literature. The research accomplished focused on three key research areas: (i) 3D objects modeling, (ii) 3D human pose estimation and (iii) 3D human motion analysis. The research performed was cross-disciplinary in nature, with exhaustive analysis in interpreting and understanding how 3D information can be gathered from 2D images in general.

In Chapter 1 a general introduction to the focal point research of the thesis is done. In Chapter 2 the main objectives of the research are outlined, providing the justification under the doctoral studies. In Chapter 3 a review of the literature regarding the key research areas is described. In Chapter **??** a general background is introduced.

The major research objectives for this research were evaluated through a series of scientific papers centred on the key research areas mentioned above. The 3D modeling of objects was achieved examining articulated objects and concave and reflective objects. The studies undertaken are described in Chapter 4 and 5. The paper *Component-wise modeling of articulated objects* outlines a framework for 3D modeling of articulated objects based on the aspects of their components. The paper *Single image object modeling based on BRDF and r-surfaces learning* outlines a framework for 3D modeling of concave objects with refective surfaces. In Chapter 6 is described the paper *Bayesian Image based 3D Pose Estimation* which introduces a method for 3D human pose estimation based on hierarchical Bayesian non-parametric models. In Chapter 7 is described the paper *Discovery and recognition of motion primitives in human activities* which introduces a framework for the automatic discovery and recognition of motion primitives in videos, used in turn to recognize human activities.

In Chapter 8 are presented conclusions and future research directions.

1.3 List Of Publications By Candidate

Peer Reviewed Published Journal Papers:

- **Sanzari, M.**, Ntouskos, V., & Pirri, F. (2019). *Discovery and recognition of motion primitives in human activities.* PloS one, 14.4: e0214499.

Peer Reviewed International Conference Papers:

- Ntouskos, V., **Sanzari, M.**, Cafaro, B., Nardi, F., Natola, F., Pirri, F., & Ruiz, M. (2015). *Component-wise modeling of articulated objects*. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2327-2335).
- Natola, F., Ntouskos, V., **Sanzari, M.**, & Pirri, F. (2015). *Bayesian non-parametric inference for manifold based MoCap representation*. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4606-4614).
- Natola, F., Ntouskos, V., Pirri, F., & **Sanzari, M.** (2016). *Single image object modeling based on BRDF and r-surfaces learning*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4414-4423).
- **Sanzari, M.**, Ntouskos, V., & Pirri, F. (2016). *Bayesian image based 3d pose estimation*. In European conference on computer vision (pp. 566-582). Springer, Cham.

Peer Reviewed International Conference Workshop Papers:

- **Sanzari, M.**, Natola, F., Nardi, F., Ntouskos, V., Qudseya, M., & Pirri, F. (2015). *Rigid tool affordance matching points of regard*. In International Conference on Intelligent Robots and Systems Workshops.
- Qodseya, M., **Sanzari, M.**, Ntouskos, V., & Pirri, F. (2016). *A3D: A device for studying gaze in 3D*. In European Conference on Computer Vision Workshops (pp. 572-588). Springer, Cham.
- Mauro, L., Alati, E., **Sanzari, M.**, Ntouskos, V., Massimiani, G., & Pirri, F. (2018). *Deep execution monitor for robot assistive tasks*. In Proceedings of the European Conference on Computer Vision Workshops.

Chapter 2

Aims and Objectives

The aim of the project is to implement a framework able to tackle human activity recognition from RGB videos, a wide problem leading to several considerations.

In order to deal with human activity recognition from videos, human motion is the most important problem to be faced. Besides human motion, other relevant cues can be retrived from video frames, such as the scene or the objects involved in the activity. This drives to face problems such as scene recognition, object recognition, 3D object modeling, 3D human pose estimation, 3D person tracking, etc..

A more precise definition of the objectives of this Ph.D. research project requires a description of limitations of current state of the art methodologies.

An important disambiguation has to be exposed before proceeding further: the one between action and activity. Actions are generally described in literature as single person movements that may be composed of multiple simple gestures organized temporally, such as walking, waving or and punching. Gestures are instead elementary movements of a body part. On the other hand, activities are described as involving two or more persons and/or objects, or a single person performing complex actions, i.e. a sequence of actions.

Human activity recognition from videos aims to recognize which human activities are taking place during a video, considering only features directly extracted from video frames. The related applications are manifold: healthcare monitoring applications, such as rehabilitation or stress monitoring, monitoring and surveillance for indoor and outdoor activities, human-machine interaction, entertainment etc.. Human activities can be classified into three main groups: single person, multiple people interaction and crowd behavior. In this thesis, among other computer vision problems, we will focus on single person activity recognition.

Human activity recognition has a long history in the computer vision research community, which is filled with a diverse number of approaches for activity description and modeling. Among the computer vision problems, activity recongition is still an hot topic. The reason for this variety and for

the growing interest is that there is no single model which is able to generalize well in every case. This requires the development of new techniques to improve the accuracy of recognition algorithms under more realistic conditions.

The main consequence is that the great majority of existing algorithms are customized for the specific activities needed to be recognized. This can lead to optimal results for a specific subset of human activities, but to very poor results when scaling up to a great number.

Exsisting approaches for activity recognition also include the usage of external and wearable devices different from cameras. Smart homes are a usual example of external sensing. These systems are able to recognize activities such as eating, taking a shower, washing dishes, etc., because they depend on data emerging from sensors placed in objects which people are presumed to interact with. Approaches including the utilization of wearable sensors relies on quantified attributes related to the typical movement of users, for example using accelerometers, or are related to environmental parameters such as temperature and humidity, or physiological signals such as breath or heart rate. Other approaches include the usage of special equipment for 3D human skeleton joints tracking, such as Vicon or X-sense. The extracted 3D joints poses are the so called MoCap (Motion Capture) data, and are obtained through wearable sensors (gyroscopes) or markers reflecting infrared lights coming from several cameras. The main problem related to these approaches is the need of user engagement, both to interact with specific object or to wear sensors or markers.

The limitations of the approaches explained so far lead to the conclusion that activity recognition from RGB cameras is the favorite solution for this problem.

A wide variety of approaches for activity recognition from RGB videos based on human features have been suggested so far, focusing on human features segmentation, extraction and representation. These approaches make use of features catching space and time relationship, the so called space-time volumes (STV), or discrete Fourier transform (DFT) of image frames, able to capture image intensity variations. The STV and DFT are global features which are retrieved considering the whole image. Other methods make use of local features insead. Local features, such as SIFT, HOG, etc., are needed to prevail problems such as noise and occlusion, and potentially to rotation and scale. In addition to global and local features, other processes are also suggested to directly or indirectly model human body, to which the 3D pose estimation and body part tracking methodologies can be applied. After selecting suitable features from frames or videos, activity detection and classification algorithms are the next step. One of the most employed classification algorithms is dynamic time warping (DTW), which is a similarity measure for two sequences. The main drawback of DTW is that it needs large templates. To overcome this issue, many model-based methods are proposed. Model-based techniques can be partitioned into generative models

and discriminative models. Generative models, that specifically reproduce the generation process of the data sequences as managed by the hidden Markov model (HMM) and dynamic Bayesian network (DBN). Discriminative models produces fewer hypothesis on the distributions but depends deeply on the quality of the data, such as support vector machines (SVMs), relevance vector machines (RVMs) and artificial neural networks (ANNs). The main problem related to this approaches is that there is no single model which is able to generalize well in every case. This approaches lead to very poor results when scaling up to a great number activities to be recognized.

Most approaches developed in the computer vision literature on activity recognition focused on examining individual motion patterns of activities as attested by popular activity datasets. These techniques model activities individually and point to learn discriminative patterns for each activity class. However, activities in usual and natural scenes hardly happen separately. The interdependence between activity classes produces important cues for activity recognition. Jointly modeling and recognizing connected activities in space and time can improve recognition accuracy.

The limitations explained so far lead to the conclusion that a robust system for activity recognition cannot rely only on human motion features, but has to be context-aware.

It has been proven in (Oliva and Torralba, 2007) that context is very important in human visual systems. As there is no official definition of context in video analysis, we can consider all the involved objects and motion regions as providing contextual information about each other, as well as the scene where the activities are taking place. The most interesting elements related to methods for activity recognition are emerging in those researches that are able to connect several aspects leading to the recognition of human activities in context (Caddigan et al., 2017; Rosenfeld and Ullman, 2016; Ramanathan et al., 2015; Cheron et al., 2015; Jiang et al., 2011). This new wave is also made possible thanks to the availability of image databases such as MSCoCo (Lin et al., 2014) and the VisualGenoma (Krishna et al., 2016) advancing the state of the art in the direction of building context knowledge (Jiang et al., 2011). Most of the exsisting approaches for activity recognition from context investigated human-object interaction, both in still images or videos, or perform contemporary human and object tracking to achieve action recognition.

The available state of the art research still misses a complete framework for human activity recognition based on context, taking into account both the scene where activities are taking place, 2D and 3D objects analysis, 3D human motion analysis and interdependence between activity classes.

To successfully recognize an activity in a video some important cues can be employed, such as the cause effect relation connecting a subject pose and some object in the scene, the time persistence of this relation during a video, the recognition of other relevant objects not manipulated, the scene where the activity is being performed. Another fundamental step to enable a robust

activity recognition framework is to consider the interdependence between activity classes. Relevant features can be computed related to objects, to subjects pose and features enabled by tracking the poses and the interactions between subjects-objects and amid relevant objects. The research conducted so far during the Ph.D. project in 3D objects modeling, 3D human pose estimation and 3D human motion analysis will enable the discovery of some such relevant features.

All the deficiencies of state-of-the-art methodologies for activity recognition brought us to explore computer vision areas such that 3D object modeling, 3D human pose estimation from single images and 3D human motion analysis and human activity recognition from RGB videos.

Our main contributions so far for context-aware activity recognition are: the 3D modeling of articulated objects, (Ntouskos et al., 2015b); an approach for understanding objects affordances studying the relation between points on the 3D object model and the points of regard of a person picking up a the object and looking at it, (Sanzari et al., 2015); a method to model complex objects such as concave or with reflective surfaces (Natola et al., 2016); a method for the recognition of human actions based on skeleton groups, (Natola et al., 2015b); a method for 3D human pose estimation from single images (Sanzari et al., 2016); and finally a method for activity recognition based on human motion primitives (Sanzari et al., 2019).

Chapter 3

Literature review

3.1 3D Object Modeling

The computer graphics, computer-aided design and computer vision literatures are filled with an amazingly diverse number of approaches to surface description. The reason for this variety is that there is no single representation of surfaces that satisfies the needs of every problem in every application area.

Among all the computer vision problems, the single image modeling is one of the most difficult amid the ill-posed inverse problems.

As a consequence, one needs to make additional assumptions on the object's geometry (such as piecewise planarity (Hoiem et al., 2005; Liebowitz et al., 1999)), its albedo (shape from texture (Malik and Rosenholtz, 1997)), its reflectance properties (shape from shading (Malik and Rosenholtz, 1997)), or the image formation process (shape from defocus (Favaro and Soatto, 2005)).

Because these techniques make strong assumptions on shape, reflectance, or exposure, they tend to produce acceptable results for only a restricted class of images.

Existing approaches using image cues were demonstrated to generate plausible 3D models, but they all impose more or less strong limitations on the applicability to objects in real-world images. Moreover, many of these approaches give rise to hard computational challenges.

For a single image modeling problem, the surface model is obtained by minimizing a surface smoothness objective function subject to constraints from the apparent contour or from other image cues.

The main ingredient of the deformation algorithm is a variational minimization problem, whose solution, given certain modeling constraints, is the desired modified surface.

Variational minimization problems deal with maximizing or minimizing functionals, often expressed as definite integrals involving functions and their derivatives. The interest is in extremal functions that make the functional attain a maximum or minimum value. The extrema of functionals may be

obtained by solving the associated Euler–Lagrange partial differential equations, specifying some initial conditions.

Because the problem is highly ill-posed, one needs other information to get a plausible solution, and which can be expressed as constraints of the modeling problem. This information is extracted from the image taken in consideration, the most important is the apparent contour of the object, and many others can be the light hitting the object, its brightness, its color.

Regarding the state of the art literature for the problem involving modeling articulated objects, the great majority of these approaches rely on kinematic, that is, the shapes are seen as compositions of geometrically or algebraically defined primitives, connected by joints. These approaches are generally computationally very expensive and require strong user input.

Pioneering work for the modeling of curved surfaces was done by Terzopoulos et al. (Terzopoulos et al., 1988b; Terzopoulos et al., 1988a) although their reconstructions are restricted to tube-like shaped objects.

Following the deformation methods introduced by Terzopoulos, shape generation from images provides good results by exploiting the contour generator.

The work of Prasad et al. (Prasad and Fitzgibbon, 2006; Prasad et al., 2010) made further significant advances towards the modeling of arbitrary curved surfaces and generalized the class of reconstructable objects to those of higher genus. In practice, however, only objects with rather simple topology can be modeled due to the use of a parametrized surface representation.

Similar to this work, Zhang et al. (Zhang et al., 2002) and Töppe et al. (Töppe et al., 2011) introduced a single view modeling method able to compute silhouette-consistent minimal surfaces from a depth map and from a user specified volume.

The single view modeling methods, however, are not suitable for modeling articulated objects since some of their assumptions become not valid. In particular, the components of the object do not share the same plane of symmetry.

Early approaches to this problem suggested a hierarchical composition of the object components, represented as generalized cylinders (Binford, 1971), geons (Biederman, 1987), or superquadrics (Dickinson et al., 1990; Pentland, 1986). In these early works, components were modeled with parametric 3D shapes of few degrees of freedom, leading to limited resemblance to the actual geometry of the component.

Recently, (Cashman and Fitzgibbon, 2013) demonstrated that for certain classes of objects, deformable models can be used to learn the shape of the object, based on the apparent contour imaged in different configurations and a rough initial model of the object.

Multiple view modeling of different object classes from few images have been successfully obtained using networks of objects with similar viewpoints (Carreira et al., 2014), or for large scale mean shape reconstruction (Vicente et al., 2014).

We have recently proposed a novel approach to modeling articulated objects from few images based on multiple views of the object's main parts, (Ntouskos et al., 2015b).

The fact that surface modeling from a single view has to deal with shading and the way materials shine and reflect the light has become clear since the works of (Nicodemus, 1965) and (Horn, 1977). Though only recently a great deal of work has been done to merge the rich information that light conveys about an object with its shape. Relevant examples are studies on specular reflection of materials and light incidence (Magda et al., 2001; Mallick et al., 2005), so as to dismiss the Lambertian hypothesis, and on how illumination and reflectance combine to influence an object shape perception (Barron and Malik, 2015) and its geometry (Oxholm and Nishino, 2012).

The concept of Bidirectional Reflectance Distribution Function (BRDF) has been largely used in the computer vision community (Romeiro and Zickler, 2010) to infer the material reflectance properties of a known object. Some approaches model objects in 3D by imposing an unknown BRDF such as in (Magda et al., 2001), where the object shape is recovered with two different methods requiring, however, multiple images of the same object. Retinex theory, (Land and McCann, 1971), has been used for separating the shading component from the reflectance one, in an image. A similar distinction is made in (Barrow and Tenenbaum, 1978) for extracting the intrinsic characteristics of surface orientation, reflectance and incident illumination, from a single image.

Very recently, we proposed a method to learn a non-parametric model of surface appearance directly from the measured BRDFs in unknown illumination environment, (Natola et al., 2016).

3.2 3D Human Pose Estimation

3D human pose estimation is the process of identifying how a human body is configured in a given scene. Vision-based approaches are often used to provide such a solution, using cameras as sensors. It is an important challenge for many computer vision applications, such as surveillance, automotive safety and behavior analysis, as well as Human Computer Interaction applications.

3D human pose estimation approaches can be classified in model based and model free methods. The first ones [11,12] are those which learn a mapping between visual appearance and human body pose, while the secon ones employ human knowledge to recover the body pose. Search space is reduced considering the human body appearance and structure.

Human pose estimation from images has been considered since the early days of computer vision and many approaches have been proposed to face this quite challenging problem. A large part of the literature has concentrated on identifying a 2D description of the pose mainly by trying to estimate the positions of the human joints in the images. Recently, attention has been

shifted to the problem of recovering the full 3D pose of a subject either from a single frame or from a video sequence. Despite this is an ill-posed problem due to the ambiguities emerging by the projection operation, the constraints induced by both human motion kinematics and dynamics have facilitated the recovery of some accurate 3D human pose estimation.

Human pose estimation (HPE) has been extensively studied during the years by considering videos, 2D images and depth data, (Liu et al., 2015; Hen and Paramesran, 2009; Poppe, 2007). There exist several open problems; among them we mention variations in human appearance, clothing and background, arbitrary camera view-point, self-occlusions and obstructed visibility, ambiguities and inconsistency in the estimated poses.

Different features can be chosen to describe the different types of data. Focusing on 2D input data, some works assume the 2D body joints locations already given (Akhter and Black, 2015a), while others extract features from silhouettes such as HOG (Dalal and Triggs, 2005), PHOG (Bosch et al., 2007), SIFT (Lowe, 1999) and shape context (Belongie et al., 2002), or dense trajectories (Zhou and De la Torre, 2014).

In detail, concerning 3D human pose estimation from videos, very recently (Zhou and De la Torre, 2014) introduced a spatio-temporal matching (STM) among 3D Motion Capture (MoCap) data and 2D feature trajectories providing the estimated camera view-point and a selected subset of tracked trajectories.

Depth cues have also been considered for 3D human pose estimation, such as in (Plagemann et al., 2010) where are introduced keypoint detectors based on saliency of depth maps.

In the last years many works have approached the estimation of the poses via deep learning as in (Li and Chan, 2014; Tompson et al., 2014; Ouyang et al., 2014; Toshev and Szegedy, 2014; Mehta et al., 2018).

Assuming that joint positions are already given in 2D with the corresponding image, (Akhter and Black, 2015a) propose to learn pose-dependent joint angle limits from a MoCap dataset, to form a prior for estimating the 3D poses, together with the camera parameters.

A novel class of descriptors, called tracklets, is defined and 3D poses are recovered from them. In (Lehrmann et al., 2013), human pose is estimated via a non-parametric Bayesian network and structure learning, considering the dependencies of body parts.

3.3 Human Action Recognition

Human action recognition is still a challenging and stimulating problem especially when considering motion capture data (MoCap), which are relevant in several applications including robotics, sports, rehabilitation and entertainment. A considerable amount of work has been proposed so far to solve problems arising in action recognition, such as view-point change, occlusions, likewise

variations in behaviors amid different subjects performing the same action. However there is a significant difference between MoCap and 2D/2.5D action representations, and it could be argued without fear that the two recognition problems are drastically different, as they address different feature spaces and representations and, consequently, different recognition methods. MoCap sequences represent actions by 3D points, and joints of the human skeleton with appropriate kinematics. These data can be acquired by means of an RGB-D sensor, such as the Kinect, by infrared marker tracking systems, such as the Vicon System, or via back-projection techniques using multiple cameras. With this kind of data, occlusions so far has not been considered a major issue, such as with 2/2.5 D data, however variations amid behaviors is still a major problem to be handled. Among the most relevant approaches we recall (Gong and Medioni, 2011; Lv and Nevatia, 2006; Harandi et al., 2014; Ofli et al., 2012; Wang et al., 2014b), all using noise and occlusion free datasets. In (Gong and Medioni, 2011) actions are represented as structured-time series, with each frame lying on a high-dimensional ambient space, from which a spatio-temporal manifold is obtained by a dimensionality reduction approach, based on dynamic manifold warping, accounting only for joints translation. In (Vemulapalli et al., 2014), instead, both joints rotations and translations are considered, so as to construct a novel class of features in $SE(3) \times \cdots \times SE(3)$, obtaining a full feature space mapped on the Lie algebra. In (Harandi et al., 2014) actions are represented via joint covariance descriptors, so as to work with symmetric positive definite matrices, which lie on Riemannian manifolds. In most of the approaches the representation of the joints space is a major issue and the need for a viable compromise between space reduction and completeness seems evident.

Indeed, the representation model is crucial, both for eliciting features and for the recognition method used. For example, (Gong and Medioni, 2011; Vemulapalli et al., 2014; Lv and Nevatia, 2006) consider a time-based ordering for which a temporal alignment is needed. In particular, (Lv and Nevatia, 2006) decompose the 3D joints into subspaces representing either the motion of a single body part, or of the combination of multiple ones.

Other approaches considering behaviors classification are (Ofli et al., 2012; Wang et al., 2014b). In (Ofli et al., 2012), the most informative joints are extracted by considering the fastest joints or the joints that mostly vary in angles. However, this approach proves to be effective only if simple actions are considered. Similarly, (Wang et al., 2014b) construct a so called "actionlet ensemble", which is a collection of the most discriminative primitive actions, which in turn are the representative features of subsets of joints of an action sequence. These actionlets are learned within support vector machine framework.

Recent advances in action recognition have been made thanks to the growing capabilities of neural networks. Convolutional Neural Networks (CNNs) and deep Convolutional Neural Networks (3D- CNNs) have been used for action

recognition from RGB or RGBD videos encoding single or multiple video frames, (Feichtenhofer et al., 2016; Yang et al., 2015; Wang et al., 2016; Karpathy et al., 2014; Simonyan and Zisserman, 2014). CNNs applied to motion features obtained from RGB videos (as for example optical flow features) have been used in (Simonyan and Zisserman, 2014; Varol et al., 2018).

Long Short-Term Memory (LSTM) networks, recurrent neural networks having the capability to process sequences, have been proposed for action recognition in (Donahue et al., 2015; Yue-Hei Ng et al., 2015; Baccouche et al., 2010; Baccouche et al., 2011) where video frames and motion features are given as input to the networks. Also Attention-LSTMs (ALSTMs), taking into account special frame regions in the form of attention, have been proposed in (Li et al., 2018).

The main problem related to neural networks applied to a set of video frames is that processing them highly increases the learning complexity. To compensate for this problem very large datasets are required.

3.4 Action Recognition Based On Human Action And Motion Primitives

While human motion in its generality is a vast research area, the paradigm of motion primitives has mainly been explored from the point of view of action primitives and temporal segmentation of actions. Many approaches have explored video sequences segmentation to align similar action behaviors (Gong et al., 2014), or for spatio-temporal annotation as in (Lillo et al., 2016). Lu et al. (Lu et al., 2015) propose to use a hierarchical Markov Random Field model to automatically segment human action boundaries in videos. Similarly, (Bouchard and Badler, 2007) develop a motion capture segmentation method. n-grams have been used to achieve action recognition based on action primitives. In (Thurau and Hlaváč, 2007; Thurau and Hlavác, 2008) activities are represented as temporal sequences of primitive poses. In (Thurau and Hlaváč, 2007) action primitives are extracted reducing the dimensionality of silhouettes binary images with principal component analysis (PCA). In (Thurau and Hlavác, 2008) pose primitives are extracted using HOG features. As a matter of fact, many of the earliest more relevant approaches share the paradigm that understanding human motion requires view independent representations and that a fine grained analysis of the motion field is paramount to identify primitives of motion. In early days this required a massive effort in visual analysis to obtain the poses, the low level features, and segmentation. Nowadays, scientific and technological advances have made it possible to exploit several methods to measure human motion, such as the availability of a number of MoCap databases (Ionescu et al., 2014; Mandery et al., 2015; Sigal et al., 2009). Furthermore, recent findings result in methods that can deliver 3D

human poses from videos if not even from single frames (Sanzari et al., 2016; Zhou et al., 2016; Tome et al., 2017).

Many research areas have taken advantage from the use of motion primitives, in robotics for learning by imitation studies (Gams et al., 2016; Pastor et al., 2009; Kulić et al., 2012) or learning task specifications (Ureche et al., 2015) where primitives are analyzed from a dynamical point of view, or represented as hidden Markov models (Inamura et al., 2004; Asfour et al., 2006; Billard et al., 2006). In Neurophysiology by (Bizzi and Mussa-Ivaldi, 1995; Flash and Hochner, 2005; Flash et al., 2013; Viviani and Flash, 1995; Flash and Handzel, 2007), where the idea that kinetic energy and muscular activity are optimized in order to conserve energy is commonly employed.

Our view on motion primitive shares this hypothesis of energy minimality during motion, likewise the idea to characterize movements using the proper geometric properties of the skeleton joints space motion. However, for primitive discovery, we go beyond these approaches capturing the variation of the velocity of a group of joints using this as the baseline for computing the change in motion by maximizing the motion flux.

Besides these works, only (Vecchio et al., 2003; Yang et al., 2013b; Holte et al., 2010; Endres et al., 2013) have targeted motion primitives, to the best of our knowledge. (Vecchio et al., 2003) focuses on 2D primitives for drawing, on the other hand (Yang et al., 2013b) does not consider 3D data and generate the motion field considering Lukas-Kanade optical flow for which Gaussian mixture models are learned. Furthermore, they do not provide quantitative results for motion primitives, but only for action primitives, which makes their method not directly comparable with ours. To the best of our knowledge, only (Holte et al., 2010) uses 3D data and explicitly mentions motion primitives, providing quantitative results. The authors also consider optical flow to account for the velocity field and focus on the recognition of motion primitives basing on harmonic motion context descriptors. In particular, primitive discovery is contextual to recognition. Since (Holte et al., 2010) deal only with upper torso gestures we compare with them only the primitives they mention. And we note, finally, that in our method start and end are unknown, and primitive discovery is modeled by motion flux.

3.5 Context-Aware Action Recognition In Videos

Several authors have investigated human-object interaction in still images. Yao et al. in (Yao and Fei-Fei, 2010b) use a random field model with structure learning methods to learn patterns of connectivity between objects and human body parts. Prest et al. in (Prest et al., 2012) define a human-object interaction model learning the probability distribution of human-object spatial relations. In (Yao and Fei-Fei, 2010a) a grouplet feature is proposed to recognize human-object interactions. Among other approaches working on context in still images

(Ramanathan et al., 2015) investigates a method to devise how actions are related to each other in order to give a finer structure to the interpretation of complex actions.

In videos human-object interaction has been addressed modeling the spatio-temporal evolution of interactions between persons and objects. (Escorcia and Niebles, 2013) model the spatio-temporal evolution of human-object interactions introducing a related descriptor. (Lea et al., 2016) introduce a model for action segmentation and classification from videos combining spatio-temporal features in a spatio-temporal CNNs, capturing changes in objects relationships during actions execution.

Visual trajectories have been investigated in (Packer et al., 2012; Prest et al., 2013). Packer et al. in (Packer et al., 2012) perform action recognition and object tracking using range and video sensors together with a real time human pose tracker, describing actions as visual trajectories and constructing a latent structural SVM to model manipulated objects. (Prest et al., 2013) perform action recognition from videos modeling actions as trajectories of objects with respect to humans positions, tracking objects and human poses over time.

In (Yang et al., 2013a) the authors propose social network analysis based features to encode relations between humans and objects to achieve human-object interaction recognition from videos.

Several researchers investigated specific approaches for videos recorded using first-person cameras, such as (Behera et al., 2012; Pirsiavash and Ramanan, 2012). In (Behera et al., 2012) the authors propose an approach for real-time egocentric recognition of activities constructing histograms of atomic events representing relationships between body parts and objects. Finally, in (Pirsiavash and Ramanan, 2012) recognition of activities of daily living from videos is performed exploring temporal structure and interactive models of objects.

Chapter 4

Component-wise modeling of articulated objects

Valsamis Ntouskos[1], Marta Sanzari[1], Bruno Cafaro[1], Federico Nardi[1], Fabrizio Natola[1], Fiora Pirri[1], & Manuel Ruiz[1]

[1] ALCOR LAB, Dipertimento di Ingegneria Informatica Automatica e Gestionale, Sapienza University of Rome

Published: In Proceedings of the IEEE International Conference on Computer Vision (pp. 2327-2335), 2015.

Statement of Contributions of Joint Authorship

Valsamis Ntouskos (Research Colleague):
Writing and compilation of manuscript, established methodology, data analysis, preparation of tables and figures, co-author of manuscript.

Marta Sanzari (Candidate):
Writing and compilation of manuscript, established methodology, data analysis, preparation of tables and figures, preparation of software code.

Bruno Cafaro (Research Colleague):
Minor co-author of manuscript.

Federico Nardi (Research Colleague):
Minor co-author of manuscript.

Fabrizio Natola (Research Colleague):
Minor co-author of manuscript.

Fiora Pirri (Principal Supervisor):

Supervised and assisted with manuscript compilation, editing and co-author of manuscript.

Manuel Ruiz (Research Colleague):
Minor co-author of manuscript.

The initial research conducted for the Ph.D. project covered 3D object modeling from few images of simple objects such as convex ones (Ntouskos et al., 2015b). The motivation behind these works related to the goal of the project, namely activity recognition based on context, is that the 3D shape of the objects involved in the scene can highlight relevant information for the recognition of object manipulability and interaction with persons.

Here we present the work done for 3D articulated objects modeling, published at the International Conference on Computer Vision 2015, which introduces a method for computing 3D models of articulated objects, by decomposing them into components and reassembling them using two or more images of the object in a reference pose. In particular, aspects of articulated objects are segmented from images downloaded from the web. 3D models of aspects are than reconstructed using the strain energy functional. Objects components, i.e. objects fundamental parts, are assembled toghether and finally, the complete object 3D model is built. Furthermore, software code for this paper was made available at `https://github.com/alcor-lab/articulated-object-modeling`.

This Chapter is an exact copy of the conference paper referred to above.

4.1 Abstract

We introduce a novel framework for modeling articulated objects based on the aspects of their components. By decomposing the object into components, we divide the problem in smaller modeling tasks. After obtaining 3D models for each component aspect by employing a shape deformation paradigm, we merge them together, forming the object components. The final model is obtained by assembling the components using an optimization scheme which fits the respective 3D models to the corresponding apparent contours in a reference pose. The results suggest that our approach can produce realistic 3D models of articulated objects in reasonable time.

4.2 Introduction

The problem of modeling articulated objects, like people, animals and complex human artifacts has a long history in computer vision. Obtaining 3D models of objects from images is essential for many high-level vision tasks. Early approaches suggested a hierarchical composition of the object components, represented as generalized cylinders (Binford, 1971), *geons* (Biederman, 1987), or superquadrics (Pentland, 1986; Dickinson et al., 1990), just to cite a few well known approaches to the structural descriptions theory. In these early works, components were modeled with parametric 3D shapes of few degrees of freedom, leading to limited resemblance to the actual geometry of the component.

With the popularization of more accurate deformable models, introduced also by the computer graphics community (see (Botsch and Sorkine, 2008) for a review), more realistic models of the components of an object can be obtained. Recent works (Prasad et al., 2010; Töppe et al., 2013; Vicente and Agapito, 2013) have successfully shown how some types of animals can be modeled from a single image, relying mainly on the symmetry of the animal's shape and, possibly, on further image cues. These approaches differ from the ones proposed in computer graphics (e.g. (Nealen et al., 2007; Chen et al., 2013; Levi and Gotsman, 2013)), since there, input from the 3D artist is essential. The single view modeling methods, however, are not suitable for modeling articulated objects since some of their assumptions become not valid. In particular, the components of the object do not share the same plane of symmetry.

In this work, we provide a solution to the problem of modeling articulated objects by explicitly modeling their components from various aspects. We consider a hierarchical decomposition of the object into components. Depending on the geometric complexity of the component, a different number of aspects is required for the modeling. For example, the body of an animal typically requires three to four representative aspects (left, right, front and back), while the legs can be modeled also by a single aspect. An example of the decomposition in components and aspects is presented in Figure 4.2. From each aspect, namely a particular view of the component, an approximate model of the imaged component is obtained using the deformation paradigm. Then, these aspect models are merged together to form a component. Components are typical of an object class and, in turn, are assembled considering a reference pose of the object, providing a 3D model of the whole object. Here, we assume that the object components are segmented out in the respective aspects. It is important to note that the different aspects need not correspond to the same physical object as soon as objects of the same class are sufficiently similar. We focus our study on animals as they typically satisfy this property. An example of a 3D model obtained with our approach is shown in Figure 4.1.

The paper is organized as follows. In the next section we review related work. In Section 4.4 we describe how components are modeled by their aspects,

Figure 4.1. Left: Images of an animal downloaded from the web, **Right:** 3D model obtained with the proposed method.

while in Section 4.5 we show how components are assembled to form the final model. In Section 6.6 we provide an evaluation of the proposed method and Section 4.7 adresses conclusions and future work.

4.3 Related work

Approaches based on geometric modeling of objects have recently become popular in computer vision. Following the deformation methods introduced in the pioneering work of Terzopoulos (Terzopoulos et al., 1988a), shape generation from images has been proved to provide good results by exploiting the contour generator of the object. Koenderink (Koenderink, 1984) establishes a general rule relating the curvature of the contour and the curvature of the surface, which is also investigated in (Cipolla and Giblin, 2000). Single view modeling of objects with predefined genus and topology has been introduced in (Prasad et al., 2006; Prasad et al., 2010) using images of the same object family. Recently, (Cashman and Fitzgibbon, 2013) demonstrated that for certain classes of objects, deformable models can be used to learn the shape of the object, based on the apparent contour imaged in different configurations and a rough initial model of the object. Additional image cues have been considered in (Oswald et al., 2012; Töppe et al., 2013) to model object classes from single views, and a similar approach has been taken by (Vicente and Agapito, 2013), exploiting the contour generator. A recent review is found in

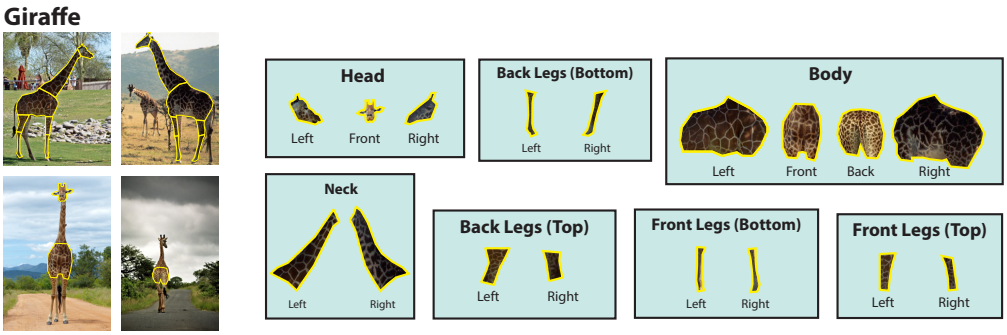


Figure 4.2. Object decomposition and aspects of each component: **Left:** Images of the object overlayed with segmentation masks, **Right:** Representative aspects of each component.

(Oswald et al., 2013).

On the other hand, the relation between the apparent contour and the contour generator, that we exploit here for reassembling the components, has been studied since the early days of computer vision. As mentioned above, Koenderink in (Koenderink, 1990) studies various properties of the contour generator based on the results of differential geometry, though a comprehensive study of the properties of contour generator of evolving implicit surfaces is found in the work of (Plantinga and Vegter, 2006). The problem of fitting 3D objects in their apparent contour has been treated in (Cashman and Fitzgibbon, 2013) where an optimization is performed to find 3D-2D correspondences, considering a parametric representation of the surface and an estimation of the view direction, initialized by the user. The problem has been also treated in (Budd et al., 2013) for non-rigid surface sequences. Our method, instead, uses a global optimization scheme, in order to estimate the view direction without requiring user input.

Finally, to improve visual quality, surface smoothing is applied at the models. Level-set based methods have been widely used for this task (for a survey see (Calakli and Taubin, 2011)). These approaches use an implicit function representation of the surface and have the advantage of topological flexibility. We follow the approach of (Liang et al., 2013), exploiting also the convenience implicit surfaces offer in performing Boolean graphics operations, for obtaining a model with no internal faces.

4.4 Modeling object aspects into components

In this section we present our approach for modeling the components of an articulated object, such as an animal, by generating models corresponding to its representative aspects and then merging them together. For each component, we assume that a set of N_c images $\{I\}_{i=1}^{N_c}$ of different representative aspects of the component are collected. Moreover, the component is segmented out in the

image, giving the corresponding aspect masks $\{A_i\}_{i=1}^{N_c}$. Note that components may share images as in shown in Figure 4.2. Moreover, the correspondence of the aspect masks with the components is given. The method is divided in two steps:

1. Given the segmentation masks of each aspect, a corresponding aspect model is generated.
2. Models of the components are built by merging together the aspect models of the component.

Figure 4.4 shows the results of these two steps for the component 'head' of the giraffe.

4.4.1 Aspect modeling

The basic idea develops on the minimization of an elastic energy that deforms the distance between nearby points, inducing local stretching and bending. Given a mask $A{\in}\mathbb{R}^2$, the surface $\boldsymbol{\varphi}{\subset}\mathbb{R}^3$ parametrized by the function $r{:}\,A{\mapsto}\mathbb{R}$ is computed by minimizing the strain energy functional defined by the first and second fundamental forms (Terzopoulos et al., 1987), plus an additional regularization term. A linearization of the energy strain is attained by considering the first and second derivatives of r (Botsch and Sorkine, 2008). The energy functional is:

$$E(r) = \int_A \mathbf{r}_{\mathcal{S}}^{\top}\mathbf{Q}_{\mathcal{S}}\mathbf{r}_{\mathcal{S}} + \mathbf{r}_{\mathcal{B}}^{\top}\mathbf{Q}_{\mathcal{B}}\mathbf{r}_{\mathcal{B}} - 2fr \;\mathrm{d}u\mathrm{d}v \tag{4.1}$$

with $\mathbf{r}_{\mathcal{S}}{=}(r_u, r_v)^{\top}$, $\mathbf{r}_{\mathcal{B}}{=}(r_{uu}, r_{vv}, r_{uv})^{\top}$, $\mathbf{Q}_{\mathcal{S}}$ is a $2{\times}2$ matrix holding the stretching parameters and $\mathbf{Q}_{\mathcal{B}}$ is a diagonal $3{\times}3$ matrix holding the bending ones. Regularization conditions are applied to make the final surface growing faster near the boundary and where the initial mask is thinner and convex. The regularization is enforced by the function $f{:}\,A{\mapsto}\mathbb{R}$ defined by:

$$\begin{aligned} f(u,v) = &[\delta_1(u,v)\gamma_1 + (1{-}\delta_1(u,v))\gamma_2]\frac{h(u,v)}{d(u,v)} + \\ &[\delta_2(u,v)\gamma_3 + (1{-}\delta_2(u,v))\gamma_4]\big(\bar{h}{-}h(u,v)\big) \end{aligned} \tag{4.2}$$

with $h(u,v){=}\mathrm{dist}\{(u,v), \partial A\}$; $d{:}\,A{\mapsto}\mathbb{R}$ is the distance between the point (u,v) and its opposite point in the normal direction, given by its nearest orthogonal projection on the boundary; $\bar{h}{=}\max h(u,v)$; $\delta_1{:}\,A{\mapsto}\{0,1\}$ is the indicator function of the convexity of a point relative to the mask; $\delta_2{:}\,A{\mapsto}\{0,1\}$ is a thresholding indicator function such that $\delta_2{=}1$ if $h(u,v){<}\tau$ and $\delta_2{=}0$ otherwise; $\gamma_i, \tau{\in}\mathbb{R}_+$, $i{=}1,2,3,4$, are the respective weights with $\gamma_1{<}\gamma_2$ and $\gamma_3{<}\gamma_4$.

The scheme for finding the solution $r(\cdot)$ of the energy functional (4.1) is based on the Finite Element method, as described in (Celniker and Gossard, 1991), applied to the associated Euler-Lagrange equation. The approximation

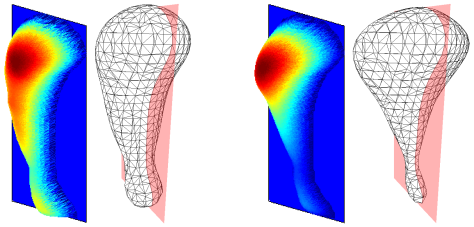


Figure 4.3. Comparison of the solutions (height maps) and reconstructed surfaces (meshes), **Left**: regularized with (4.2), **Right**: without regularization. (Best seen in colors)

of the displacement $\mathbf{r}(u, v)$ which minimizes the energy functional (4.1) is obtained as:

$$\mathbf{r}(u, v) = \mathbf{X}^\top \mathbf{\Phi}(u, v), \tag{4.3}$$

where Φ contains the coefficients of continuous shape functions and $\mathbf{X}$ unknown weights. These weights are obtained by solving the following quadratic minimization problem

$$\min_{\mathbf{X}} \left\{ \mathbf{X}^\top \mathbf{K} \mathbf{X} - \mathbf{F}^\top \mathbf{X} \right\}, \tag{4.4}$$

with $\mathbf{K}$ the stiffness matrix and $\mathbf{F}$ the regularization term. Algorithm 1 describes the steps involved. In order to constrain the solution at the boundary ∂A, homogeneous Dirichlet conditions are applied to the whole boundary of the mask into the PDE problem formulation. Once the solution is computed for the i-th aspect, $i \in \{1,\ldots,N_c\}$, the corresponding mesh B_i is obtained from the surface defined by the composition of $\boldsymbol{\varphi}$ with its reflection along the $z=0$ plane, as shown in Figure 4.3.

4.4.2 Component building

In order to obtain a model for each component, it is necessary to combine together the models $\{B_i\}_{i=1}^{N_c}$ produced from each aspect. To achieve this, 3D transformations $\{T_i^r\}_{i=1}^{N_c}$ are estimated between each aspect model B_i and a reference model B_r, in a way that produces a consistent model. Using feature points extracted from the images of each aspect I_i (see Figure 4.2), an initial registration of the models is achieved which is then refined by dense 2.5D registration. Algorithm 2 shows the steps for registering the aspects.
The last step of Algorithm 2 (line 11) is a dense 2.5D registration between the depth image of the reference aspect d_r and the depth image $d_i^{(0)}$ corresponding to the transformed i-th aspect. The registration is obtained via the following minimization problem

$$\min_{\xi_i \in \mathfrak{a}(3)} \|d_r - d_i^{(0)}(\xi_i)\|_{L_1}, \tag{4.5}$$

with $\mathfrak{a}(3)$ the Lie algebra of the 3D affine transformation group.

Algorithm 1: Aspect modeling

Input: Aspect masks $\mathcal{A} = \{A_i\}_{i=1}^{N_c}$, parameters $\{\mathbf{Q}_\mathcal{S}\}_i^{N_c}$, $\{\mathbf{Q}_\mathcal{B}\}_i^{N_c}$
Output: $\{B\}_{i=1}^{N_c}$

for A_i *in* $\mathcal{A}$ **do**
- Define a triangulation $\mathcal{T} = \{T_j\}_{j=1}^{m}$ over the points of the aspect's mask
- Choose the set of shape functions to use (e.g. linear, quadratic) and the quadrature nodes
- **for** T_j *in* $\mathcal{T}$ **do**
  - Interpolate the shape functions at the quadrature nodes
  - Assemble the stiffness matrix and regularization term using a quadrature rule
- Find the weights of the shape functions solving the equation $\mathbf{KX} = \mathbf{F}$
- Find the displacements r_i using eq. (4.3)
- Compute a mesh B_i based on the triangulation

Algorithm 2: Aspect registration

Input: Index r, $\mathcal{B} = \{B_i\}_{i=1}^{N_c}$, $\mathcal{I} = \{I_i\}_{i=1}^{N_c}$
Output: $\{T_i^r\}_{i=1}^{N_c}$

for I_i *in* $\mathcal{I}$ **do**
- Detect a set of feature points $\{F_{ij}\}_{j=1}^{M_i}$ inside the segmentation mask (e.g. by keypoints, SURF (Bay et al., 2006) features or similar)
- Project $\{F_{ij}\}_{j=1}^{M_i}$ on $\{B_i\}_{i=1}^{N_c}$ to obtain the 3D feature points $\{X_{ij}\}_{j=1}^{M_i}$

for B_i *in* $\mathcal{B} \setminus B_r$ **do**
- Find feature matches $\{F_{ij}\}_{j=1}^{M_i} \leftrightarrow \{F_{rj}\}_{j=1}^{M_r}$
- **if** #matches ≥ 3 **then**
  - Estimate 3D transformation $T_i^{(0)}$ based on $\{X_{ij}\}_{j=1}^{M_i} \leftrightarrow \{X_{rj}\}_{j=1}^{M_r}$ (up to affine transformation when the number of matches is sufficient)
- **else**
  - Ask user for manual initialization
- Apply $T_i^{(0)}$ on B_i and compute depth image $\bar{d}_i^{(0)}$
- Perform dense 2.5D registration of $\bar{d}_i^{(0)}$ w.r.t. $\bar{d}_r$

The objective function is non-smooth and non-linear in ξ. The use of L_1-norm is preferred, however, as it makes the method more robust with respect to L_2-norm. To overcome the problem of non-smoothness the Legendre-Frenchel transform is applied leading to an equivalent saddle-point problem. The objective function is then linearized with respect to ξ and a solution is computed using the nonlinear conjugate gradient method in a coarse-to-fine framework. The saddle point problem at the k-th iteration is

$$\max_{\mathbf{p}\in P} \min_{\delta\xi_i^{(k)}\in\mathfrak{a}(3)} \mathbf{p}^\top \left((\bar{d}_r - \bar{d}_i^{(k)}) - J_{\xi_i^{(k)}}\, \delta\xi_i^{(k)} \right), \tag{4.6}$$

with $\delta\xi_i^{(k)}$ the correction of ξ_i at k-th iteration, P the union of pointwise L_1 balls, $\bar{d}_r$ the discrete reference depth image, $\bar{d}_i^{(k)}$ the (discrete) depth image of aspect i transformed according to $T^{(k)} = \exp(\delta\xi_i^{(k)})T^{(k-1)}$ and $J_{\xi_i^{(k)}}$ the Jacobian of $d_i^{(k)}$ with respect to $\xi_i^{(k)}$.

The optimization significantly improves the registration provided that the initialization $\bar{d}_i^{(0)}$ is situated in the convex basin of the optimal solution. More details regarding the optimization problem (4.6) and how a solution is computed are provided in the supplementary material. The final solution depends on the choice of the reference aspect and the order in which the remaining aspects are considered, however, given that N_c is typically a small number, the solutions are equivalent.

After computing the transformations which lead to a consistent registration of the aspect models, we need to merge them into a single component model. To achieve this, we first compute a volumetric representation of each model's surface which facilitates the extraction of the surface envelope. We use the definition of Inner Product Field (IPF), as described in (Liang et al., 2013). Considering a surface S as the boundary between its interior and exterior regions, then we can represent S by computing an IPF on a regular 3D grid in the following way. Given a domain $W \subset \mathbb{R}^3$ and a set of points $S^* = \{p_i\}_{i=1}^{N_c}$ sampled from the surface, for a generic point $x \in W$ we denote by $p(x) \in S^*$ the nearest point in S^* to x. The IPF is defined as

$$\phi(x) = v(x)^\top \mathbf{n}(p(x)), \quad v(x) = \frac{x-p}{\|x-p\|}, \tag{4.7}$$

where $\mathbf{n}(\cdot)$ is the surface normal at point p. Once IPFs are computed, we have an implicit representation of the aspect surfaces and we can exploit the following result: given two or more implicit surfaces $\phi_1(x), \ldots, \phi_n(x)$, then $\phi_\cup(x) = \min(\phi_1(x), \ldots, \phi_n(x))$ is the union of their interior regions and corresponds to the envelope of the surfaces. As a final step, the component model is slightly smoothed in order to attenuate possible irregularities and artifacts. The smoothing is applied on the volumetric representation of the object using the Level Set method according to the mean curvature flow (Osher

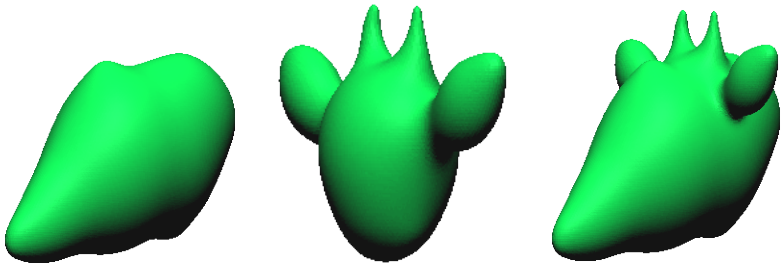


Figure 4.4. Aspects modeling and component building of the giraffe head. **Left**: side aspect, **Center**: front aspect, **Right**: component model.

and Fedkiw, 2003)

$$\phi_t + V_n \|\nabla \phi\| = 0, \tag{4.8}$$

where $V_n = -b\kappa$ is the velocity field in the normal direction generated from surface curvature κ. A mesh is then extracted by standard meshing techniques (e.g. (Lorensen and Cline, 1987)).

4.5 Assembling of the articulated object

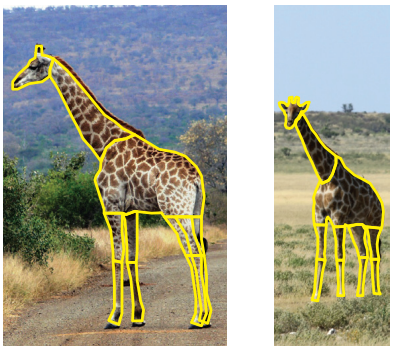


Figure 4.5. Two views of a giraffe in a reference pose with the overlayed component masks.

The components are assembled in order to reconstruct the entire object in a reference pose. In particular we use the apparent contours of the components in two or more views of the object in a reference pose, as the ones displayed in Figure 4.5. We assume here that the components are at least partially visible in these images, that the corresponding masks are available and that they are produced by an orthographic projection. The visibility requirement can be relaxed as the number of views increases.

First, we recover the optimal transformation for each component, which makes its projection comply with the apparent contour. We treat this as a 3D-2D registration problem. In particular, we consider that each component is a sufficiently smooth surface S (e.g. of class C^2) and the apparent contour is

a planar contour γ. These two entities are related by the contour generator (CG) which is a space curve Γ defined by the set of visible points on S where the view direction $\mathbf{v}$ is locally tangent. Hence, the projection of Γ according to $\mathbf{v}$ produces γ up to a 2D similarity transformation. In order to register each 3D component in its apparent contour we find a view direction and the corresponding CG which, under projection, gives a contour $\hat{\gamma}$ as similar as possible to γ.

Let $\mathcal{Y}(S)$ a set of points sampled on S. Under the given assumptions, it suffices to identify two points $Y_1, Y_2 \in \mathcal{Y}(S)$ lying on Γ, to compute the view direction. This can be seen by observing that Γ depends only on $\mathbf{v}$, and two points with non parallel normals $\mathbf{n}(Y_1)$ and $\mathbf{n}(Y_2)$ define the view direction up to a sign, as $\mathbf{v} = \mathbf{n}(Y_1) \times \mathbf{n}(Y_2)$.

We consider a discrete optimization problem, based on the energy function

$$E(Y_1, Y_2) = \sum_{l=\{1,2\}} (E_{vis}(Y_l) + E_{curv}(Y_l)) + E_{ang}(Y_1, Y_2) + E_{dist}(Y_1, Y_2). \tag{4.9}$$

The first term of (4.9) corresponds to a visibility constraint, expressing the fact that both points must be visible from the estimated viewpoint while the last three terms take into account local geometric properties that the contour and CG have to satisfy. All these terms are invariant with respect to 2D similarity transformation, which is a computational bottleneck when considered. We examine now in detail each term.

E_{curv} is based on the relation between the curvature of the surface and the curvature of the apparent contour. The curvature of γ, $\kappa^{\gamma}(\mathbf{y})$ and the curvature of Γ at the corresponding point $\kappa^{\Gamma}(Y)$ satisfy the relation

$$\kappa^{\Gamma}(Y) = \sin^2\theta\, \kappa^{\gamma}(\mathbf{y}), \tag{4.10}$$

with θ the angle between $\mathbf{v}$ and the CG at Y (Koenderink, 1990; Cipolla, 1998). Based on this result, suitable bounds regarding the curvature of γ, Γ and S are provided by the following proposition:

Proposition. *Let S be a smooth surface. Denoting as $\pi(\cdot)$ the projection operation, the curvature of the contour γ at a non-cusp point $\mathbf{y}$, the curvature of Γ at the corresponding point Y and the principle curvatures of the surface κ_1^S (minimum) and κ_2^S (maximum) at Y satisfy the inequality*

$$\kappa_1^S(Y) \leq \kappa^{\Gamma}(Y) \leq \kappa^{\gamma}(\mathbf{y}) \leq \kappa_2^S(Y), \quad \textit{with: } \mathbf{y} \in \gamma,\, Y \in \Gamma, \mathbf{y} = \pi(Y). \tag{4.11}$$

Proof. Consider a generic point $Y \in \Gamma$. We assume first that Y is not umbilical. The leftmost inequality is trivial as the curvature of Γ at Y, cannot be smaller than the minimum curvature of the surface at Y. The second inequality follows

from (4.10). To show the last inequality we consider the osculating sphere O_Y of the surface at Y which has curvature $\kappa^{O_Y} = \kappa_2^S(Y)$. Regardless of the view direction, γ at $\mathbf{y}$ can at most locally lie on the projected contour of O_Y which is a circle with curvature κ^{O_Y}. Hence, the curvature of γ at $\mathbf{y} = \pi(Y)$ is locally bounded by the curvature κ^{O_Y} which is equal to $\kappa_2^S(Y)$. If the point is umbilical then all equalities trivially hold. □

Moreover, the sign of $\kappa^\gamma(\mathbf{y})$ should match the sign of the Gaussian curvature of S at Y (Koenderink, 1990).

Corollary. *Considering a point $\mathbf{y} \in \gamma$, a region $R \subseteq S$ is an admissible region of the corresponding point $Y \in \Gamma$ iff $\kappa_1^S(\mathbf{Z}) \leq \kappa^\gamma(\mathbf{y}) \leq \kappa_2^S(\mathbf{Z})$, $\forall \mathbf{Z} \in R$ and the sign of $\kappa^\gamma(\mathbf{y})$ matches the sign of the Gaussian curvature G^S in R.*

In the following for brevity we omit the explicit relation with the surface/curve points. Based on the previous result the curvature term can be expressed as

$$E_{curv} = \sigma_\kappa^{-2}\, D_{[\kappa_1^S, \kappa_2^S]}\left(\kappa^\gamma\right) + \sigma_G^{-2} \max(-\operatorname{sgn}(G^S \kappa^\gamma), 0), \tag{4.12}$$

with $D_{\mathcal{J}}(v) = \min_{w \in \mathcal{J}}(\|v - w\|)$.

The term E_{ang} expresses the fact that the angle between the normals $\mathbf{n}(Y_1)$, $\mathbf{n}(Y_2)$ should match the corresponding angle on the apparent contour. The same holds for the angle between each of the normals and the connecting segment $(Y_2 - Y_1)$ projected on the plane spanned by the normals. This gives

$$E_{ang} = \sigma_n^{-2} c(\theta_n, \theta_\eta) + \sigma_b^{-2} c(\theta_B, \theta_b), \tag{4.13}$$

with θ_n, θ_η the angles between the 3D and 2D normals respectively, and θ_B, θ_b the angles between the base segment and one of the normals in 3D and 2D respectively. The cost function c penalizes differences between the corresponding angles (e.g. $c(\theta, \phi) = \tan(|\theta - \phi|)$). The term E_{dist} directs the search towards points which have a distance similar to the distance between the corresponding contour points. The distances are normalized with respect to the diagonal length of the corresponding entity's bounding box $d(\cdot)$, giving

$$E_{dist} = \sigma_d^{-2} \left(\frac{\|Y_1 - Y_2\|}{d(S)} - \frac{\|\mathbf{y}_1 - \mathbf{y}_2\|}{d(\gamma)} \right)^2. \tag{4.14}$$

Finally, the term E_{vis} represents the constraint that the points of the CG should not be occluded and is taken equal to the maximum penetration depth of the view ray passing through X with respect to S.

We find the global minimum of the energy function by using a branch-and-bound search strategy. First, we find the two points on γ which result into

the most restricted region on S based on the previous corollary, and use them as initial points for the search. The pair of points which corresponded to the lowest energy value gives us the view direction $\mathbf{v}$. The remaining 2D similarity transformation is then recovered by applying a shape matching technique between the resulting contour and the measured one (see (Dryden and Mardia, 1998)).

This procedure gives the relative pose of each component with respect to the view. Not depending on all the points of the apparent contour, it is robust with respect to the visible portion of the contour and the shape of the 3D component. The solution can be refined by performing an iterative LSE minimization.

By registering each component in the given view we recover their relative position with the only exception of the translation in the viewing direction. We solve this ambiguity by using the other views. In particular since the object is imaged in the same pose from two or more known views, the depth ambiguity is resolved. A single model is computed from the assembled components by following the steps presented at the end of Section 4.4.2.

4.6 Evaluation

4.6.1 Modeling time

Our implementation of the proposed method consists of a mixture of Matlab and CUDA code. The parts of our approach which allowed for massively parallel implementation were realized in CUDA. These parts are: the 2.5D registration of the modeled aspects, the computation of the IPF of each model and the surface smoothing. The parts of aspect modeling and component assembling are implemented in Matlab. A report of the time required for computing the models presented in this section is presented in Table 4.1.

Model	**AM [sec]**	**CB [sec]**	**CA [sec]**	**Sm [sec]**	**Total [sec]**
Cow	512	2.3	1125	0.10	1639
Horse	437	1.7	1451	0.05	1890
Dog	461	1.9	951.4	0.07	1414
Cat	495	2.0	1753	0.09	2250
Sheep	383	1.7	1355	0.07	1740
Hippo	501	1.8	1398	0.08	1901
Donkey	399	2.1	1224	0.10	1625
Giraffe	422	2.2	1286	0.05	1710

Table 4.1. Modeling time report (Legend: AM-aspect modeling, CB-component building, CA-component assembling, Sm-smoothing).

The experiments were performed on a Desktop PC equipped with an Intel

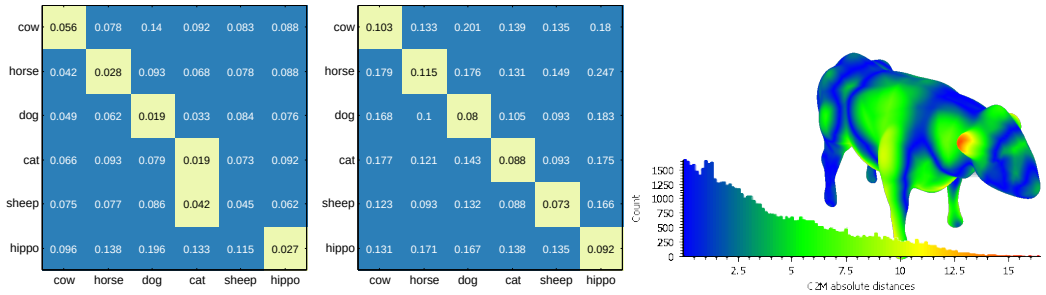


Figure 4.6. Model comparison (smallest values are highlighted), **Left:** Normalized symmetric differences between the models, **Center:** Hausdorff distances between the models, **Right:** Example of Hausdorff distance visualization for class 'cow'.

i7 3.6GHz CPU, 16GB RAM and an NVIDIA GTX970 graphics card.

4.6.2 Model comparison

We performed an extensive comparison of the models obtained with our method with respect to models downloaded from the web. Most of the downloaded models were taken from the 3D warehouse of SketchUp, while the rest were taken from other repositories. We evaluated the similarity of our models with respect to the downloaded ones using two different similarity measures, the Hausdorff distance (Aspert et al., 2002) and the normalized symmetric difference. We considered our model as reference and preprocessed the models taken from web in order to make the results comparable. The preprocessing consisted of the following steps:

i model clean-up; remove internal faces, recover manifoldness and close holes
ii manual orientation w.r.t. reference model
iii automatic non-isotropic scaling for matching the bounding box with the reference model

The Hausdorff distance was computed directly on the meshes of the models. For the symmetric difference, a volumetric representation was obtained via the IPF of the models and the distance was taken as the difference between the number of voxels that fell in the union and the number of those that fell in the intersection of the two volumes, normalized by the total number of voxels. The results of the comparison are presented in Figure 4.6 together with an example of the visualization of the Hausdorff distance on one of our models, in relation to a downloaded model of the same class. The numbers reported correspond to average values of the distances with respect to all the downloaded models of each class (3-4 models).

The results show that the models computed with our method actually represent the modeled class, as the average distance with respect to the downloaded models of the same class is consistently smaller in comparison to the distances with respect to the other classes. Further analysis shows that

the Hausdorff distance is usually higher around the parts which vary the most, as for example the neck and the belly of the animal. More images regarding the comparison between the downloaded and our models are provided in the supplementary material and the accompanying video.

4.6.3 Perceptual study

Because of the nature of the problem, similarity distances may not always be representative. To further evaluate the quality of our models we performed a perceptual study with the help of volunteers.

Experiment Ten volunteers who did not know the purpose of the study participated in the experiment. Six participants were male and four female, while 60% had from 22 to 25 years and 40% from 25 to 29 years. Finally, three subjects reported corrected-to-normal vision and the rest normal vision.

The models given in Figure 4.7 were used for conducting the study. On the left column, the models obtained with our approach are presented, while the right column contains the models downloaded from the web. The downloaded models were hand picked in order to be in a pose similar to our models' pose. Participants were invited to ask questions before the experiment. After providing the necessary information and consent the task was resented to the participants:

> "Various 3D models will be shown on the screen during the experiment. For each model, you need to identify the corresponding animal and give a mark for its quality. You can interact with the model for as long as you prefer before answering."

The models were presented on the screen with a uniform green shaded material on blue background, as shown in Figure 4.7. The participants marked the answers on a special form, where the animal class could be specified freely and a scale of discrete values from 0 to 5 was used for evaluating the quality of the model. The order models were presented was randomized to avoid bias caused by consistent ordering.

Outcome We consider the null-hypothesis H_0 that participants randomly selected the animal class, while the alternative hypothesis H_1 is that users correctly recognized the animal. Cross-tabulation was performed on the answers provided by the participants regarding the class of animal represented by our models and the resulting confusion matrix is shown in Figure 4.8. One can observe that the participants almost always identified successfully the animal class. In fact, the null hypothesis is rejected as the chi-square value is $\chi^2 = 247$, corresponding to a practically vanishing p-value. It is important to note that the participants did not know in advance the classes of animals involved. This

Cow	Horse	Dog	Cat	Sheep	Hippo
70%	50%	70%	30%	50%	50%

Table 4.2. Per-class percentage of votes above 3 (good) given to the models reconstructed by our method.

justifies also the last row of the confusion matrix, as one participant recognized the hippo as a pig.

The distribution of votes given by the participants for the model quality is presented in Figure 4.9. The hand-made models received higher votes in average, with a difference of 1.9 scale units with respect to the average vote that our models received. This is understandable considering that our models correspond to more abstract class models, lacking particular details like eyes, nose and tail. Nevertheless, the percentage of the participants which gave a vote above 3 (good) for the quality of our models (Table 4.2), indicates that the models are of satisfying quality.

4.7 Conclusions and future work

We propose a method for computing 3D models of articulated objects, by decomposing them into components. Realistic models of the object components are built by merging together 3D models obtained from different aspects, considering a type of aspect graph (Dickinson et al., 1990) which indicates the essential aspects. Aspects are extracted from images downloaded from the web. The entire object is obtained by reassembling the components using two or more images of the object in a reference pose. Our experiments suggest that our method is able to provide realistic models of the objects, both in terms of a perceptual analysis, and by a quantitative analysis of their similarity with respect to human created 3D models.

Extensions of this work in various directions can be pursued. An important extension is the possibility to model the object in different configurations by using the modeled components. This is already possible in our framework but it can be facilitated by assembling the components using a single image. This can be made possible by learning spatial relations between the components (joints, joint range etc.) and possibly also a distribution of the object poses, which would allow to compute realistic models even when some of the components are occluded. Finally, an important extension would be the automatic selection of the most representative aspects for each component from a set of images.

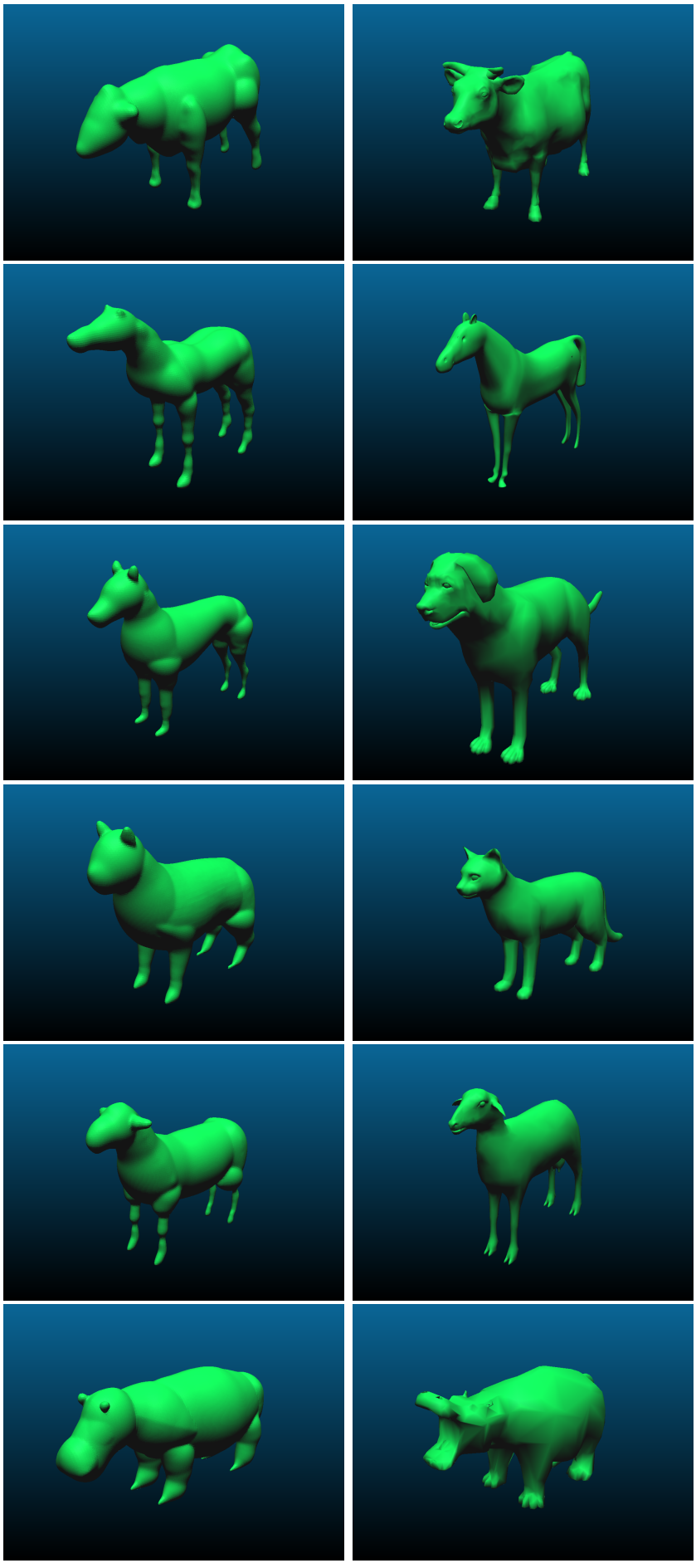


Figure 4.7. Animal models used in the perceptual study,
Left: Models computed with our method, **Right:** Models downloaded from the

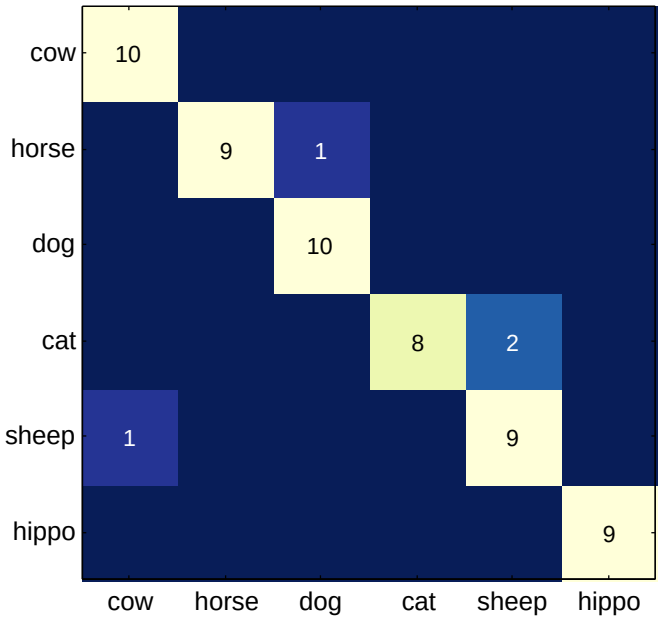


Figure 4.8. Confusion matrix from the perceptual study.

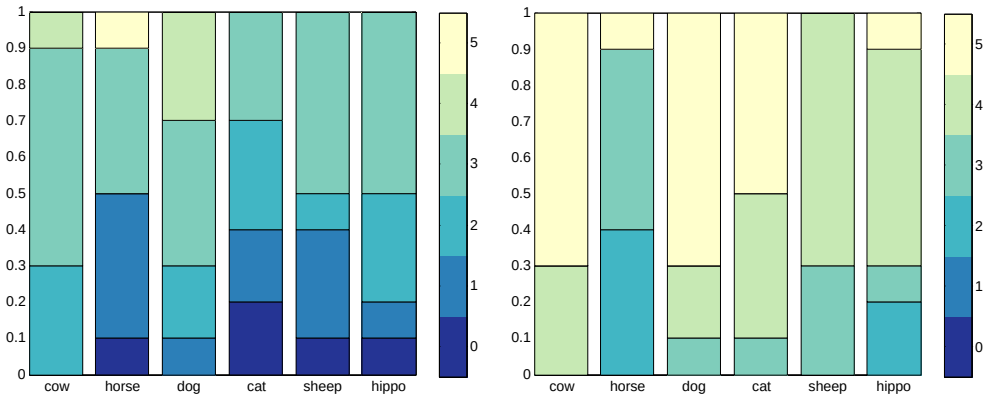


Figure 4.9. Vote distribution for the models produced with our approach (**left**) and models taken from the web (**right**).

Chapter 5

Single image object modeling based on BRDF and r-surfaces learning

Fabrizio Natola[1], Valsamis Ntouskos[1], Fiora Pirri[1], & Marta Sanzari[1]

[1] ALCOR LAB, Dipertimento di Ingegneria Informatica Automatica e Gestionale, Sapienza University of Rome

Published: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4414-4423), 2016.

Statement of Contributions of Joint Authorship

Fabrizio Natola (Research Colleague):
Editing, preparation of tables and figures, co-author of manuscript.

Valsamis Ntouskos (Research Colleague):
Writing and compilation of manuscript, established methodology, data analysis, preparation of tables and figures, co-author of manuscript.

Fiora Pirri (Principal Supervisor):
Supervised and assisted with manuscript compilation, editing and co-author of manuscript.

Marta Sanzari (Candidate):
Writing and compilation of manuscript, established methodology, data analysis, preparation of tables and figures.

Here we present the work done for 3D modeling of concave objects with reflective surfaces from single images, published at the Conference on Computer Vision and Pattern Recognition 2016, which introduces a method to model non-Lamberitan surfaces with either concave or sharp parts. We have created a synthetic dataset using 3D models of a number of real objects, obtained from different databases of 3D objects. From images obtained rendering these objects with different materials, we extracted patcheas to learn a dictionary describing the normal field for each patch. To obtain the 3D model of objects in new unseen images, we used normals and curvatures inferred using the learned dictionary. To resolve irregularities of the surface due to noise and outliers we refine the 3D surface using a photo-consistency error.

This Chapter is an exact copy of the conference paper referred to above.

5.1 Abstract

A methodology for 3D surface modeling from a single image is proposed. The principal novelty is concave and specular surface modeling without any externally imposed prior. The main idea of the method is to use BRDFs and generated rendered surfaces, to transfer the normal field, computed for the generated samples, to the unknown surface. The transferred information is adequate to blow and sculpt the segmented image mask in to a bas-relief of the object. The object surface is further refined basing on a photo-consistency formulation that relates for error minimization the original image and the modeled object.

5.2 Introduction

There is an increasing need for 3D models of objects, from single images, for several applications such as digital archives of heritage and monuments, anatomy models for pathology detection, small artifacts models for populating rendered 3D scenes with objects or augmenting a MOCAP sequence with tools for manipulation and, finally, for robotics. Likewise, there is a growing awareness that 3D modeling, from a single image, helps to navigate the sea of terabytes of images, for the object recognition challenge.

That surface modeling from a single view has to deal with shading and the way materials shine and reflect the light has become clear since the works of (Nicodemus, 1965) and (Horn, 1977). Though only recently a great deal of work has been done to merge the rich information that light conveys about an

Figure 5.1. An example of 3D surface of an object from ImageNet

object with its shape. Relevant examples are studies on specular reflection of materials and light incidence (Magda et al., 2001; Mallick et al., 2005), so as to dismiss the Lambertian hypothesis, and on how illumination and reflectance combine to influence an object shape perception (Barron and Malik, 2015) and its geometry (Oxholm and Nishino, 2012).

Here, we address these problems introducing a novel method, which is unbiased to the changes of the ambient light, taking care of both concavities and sharp parts of an object, this is the main contribution of this paper. Our approach is related to SIRFS (Barron and Malik, 2015), who introduced priors for shape, albedo and illumination, respectively, so as to learn the most likely shape. Though here we do not introduce any prior, instead we formulate an hypothesis.

Our hypothesis is that a sufficiently large number of patches, with varying surface curvature, rendered with different materials, with known reflectance properties, and varying incidence and reflection angles, can be used to estimate these properties in unknown objects. Through this generalization, the reflected, specular and diffuse light of a new object, seen in a single image, can be recovered. We show that this hypothesis is plausible and proves to give interesting results. Indeed, the normal field of the rendered surfaces, applied as an external deformation force, basing on finite element method (Strang and Fix, 1973), is used to sculpt the unknown object surface. This gives very beautiful results, that are further refined to meet photo-consistency requirements.

The paper is organized as follows. In the next section we give some pointers to related works, despite we are not able to cover the whole extraordinary literature on the topic. In Section 5.4 we introduce the basic concepts supporting the paper, namely the BRDF (Nicodemus, 1965), the MERL database (Matusik et al., 2003), how rendered surfaces (r-surfaces) are generated, and few hints for the reference database ImageNet (Deng et al., 2009) and for recovering the object contour (Vese and Chan, 2002). In Section 5.5 we introduce the unsupervised learning method to validate the hypothesis that the r-surfaces convey sufficient information about unseen objects. The distribution of the

data is inferred via a nested Dirichlet process mixture model (Ferguson, 1973; Blei et al., 2010). Features of the highest level in the hierarchy are obtained by sparse stacked autoencoders (Munro and Zipser, 1989; Olshausen and Field, 1997). The outcome is a selection of a BRDF and of the most plausible normals on each patch covering the object image. These data, as described in Section 5.6, form the external forces of the energy, which deforms the planar patches, covering the object mask, into the object surface. This extends the deformation method (Terzopoulos et al., 1987) to concavities and sharp object parts. Finally, the resulting surface model is made consistent with the object appearance in the image, by revising the light effects, as described in Section 5.7. This is obtained with a rich energy term taking care of both photo-consistency and surface depth, optimized via total variation minimization. The high level ideas of the approach are visualized in Figure 5.2. Results, shown in Section 5.8 are very promising and new, with respect to the state of the art.

5.3 Related Works

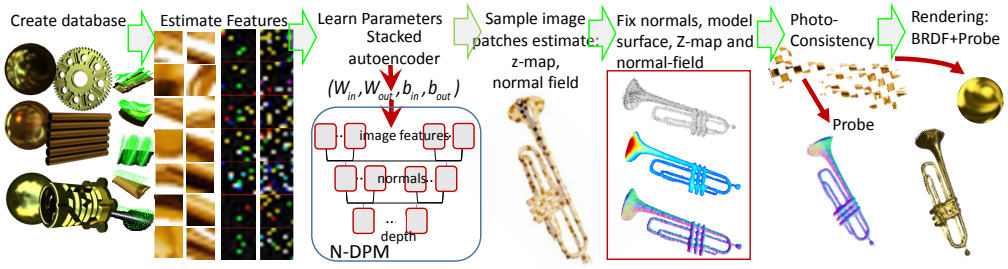


Figure 5.2. High level ideas of the work.

The concept of Bidirectional Reflectance Distribution Function (BRDF) has been largely used in the computer vision community (Romeiro and Zickler, 2010) to infer the material reflectance properties of a known object. Some approaches model objects in 3D by imposing an unknown BRDF such as in (Magda et al., 2001), where the object shape is recovered with two different methods requiring, however, multiple images of the same object. Retinex theory, (Land and McCann, 1971), has been used for separating the shading component from the reflectance one, in an image. A similar distinction is made in (Barrow and Tenenbaum, 1978) for extracting the intrinsic characteristics of surface orientation, reflectance and incident illumination, from a single image. Very recently, in (Narihira et al., 2015) the authors propose a convolutional neural network approach to separate the albedo component from the shading. Shape from Shading (SFS) recovers the shape of an object from a single image, provided the illumination and the reflectance are given, see (Zhang et al., 1999) and references therein. SFS makes strict assumptions, usually a

Lambertian material with a single light, to find the solution for the otherwise unconstrained problem. In (Oxholm and Nishino, 2012), reflectance and geometry are jointly recovered by assuming a statistical BRDF model and known lighting environment. In our work, instead, we learn a non-parametric model of surface appearance directly from the measured BRDFs in unknown illumination environment. (Richter and Roth, 2015) propose a discriminative learning approach for the SFS problem, considering an uncalibrated illumination without the assumption of a single point light. (Xiong et al., 2015) examine the light locally on small patches in a Lambertian setting and for each image patch a set of 3D surface patches, that may have generated the imaged ones, is sampled. Differently from them, our approach is not based on Lambertian assumptions. In (Saxena et al., 2009), a 3D model from a single image is reconstructed basing on super-pixels segmentation and the Random Markov Field approach. In (Chandraker et al., 2005), both inter-reflections and photometric stereo are combined to resolve the generalized bas-relief ambiguity, but in a Lambertian setting. Finally, (Vasilyev et al., 2008) consider specular objects estimating the corresponding 3D shapes by means of shape from specular flow approach with general motion.

5.4 Reflectance model and r-surfaces

In this section, we introduce some preliminary concepts concerning the BRDF, the method for rendering object surfaces (r-surfaces), and finally the segmentation algorithm for objects taken from ImageNet.

BRDF. The model considers incident directions (ϕ_i, φ_i), in spherical coordinates, defined on the local reference frame of the surface element, within some solid angle $d\omega_i$ and the direction of reflection (ϕ_r, φ_r) over some solid angle $d\omega_r$. We assume that the observer line of sight is orthogonal to the image plane and centered on the object center of mass. We assume also a geometric optics model, that is, the electromagnetic character of light can be ignored (Nayar et al., 1991). Under this hypothesis waves interference and diffraction can be disregarded. We consider three kinds of reflections: specular, diffuse, and ambient. Specular reflection, in its ideal form, is a Dirac delta function, so that $\phi_r = \phi_i$ and $\varphi_r = \varphi_i + \pi$. The specular reflection preserves the solid angle of the incident ray, namely $d\omega_i = d\omega_r$. Diffuse scattering is Lambertian, not depending on the direction of reflection. Ambient scattering collects all other kinds of reflection. In particular, lighting due to environment reflections on the surfaces is here treated as noise, so that we actually model arbitrary environment light probes.

Given the incoming light direction $d\omega_i$ and the reflected light direction $d\omega_r$, both defined with respect to the normal of an infinitesimal surface element, the BRDF (Nicodemus, 1965) is the ratio between the amount of light reflected from the surface along $d\omega_r$, namely radiance, L_r, and the total amount of light

incoming to the surface element along $d\omega_i$, namely irradiance $\mathcal{E}_i$.

There are two main databases for the BRDF values of several materials under different light conditions, the MERL Database (Matusik et al., 2003), for isotropic materials, and the UTIA one for the anisotropic materials (Filip and Vávra, 2014). We have considered the isotropic BRDFs (see (Filip and Vávra, 2014) for a discussion on isotropic and anisotropic BRDF), where the material reflectance properties are invariant under rotation of the surface about its normal. This because the MERL database is rich of most of the everyday objects materials like aluminum, brass, chrome, plastic, and acrylic.

3D models and surface rendering. We have created a synthetic dataset using 3D models of a number of real objects, obtained from different databases such as 3D Warehouse and TurboSquid. To ensure a wide variety of surface curvatures and curvature maps in our dataset, and to guarantee its semi-completeness, we consider a number S of both smooth objects, such as tubes and rings, and irregular ones such as gear wheels, see Figure 5.2, Panel 1, for some examples. Each object surface is then rendered with Blender. Each of the obtained r-surfaces, is of dimension $m \times m$ pixels, with $m \in \{256, 512\}$ and, such that for each angle of incident and reflected light $(\phi_i, \varphi_i, \phi_r, \varphi_r)$, and BRDF material, an r-surface is made available. Note that the light direction varies according to (ϕ_i, φ_i), while the view direction according to (ϕ_r, φ_r). Light is distributed considering a hemisphere with the surface at the center of it. The angles ϕ_i and ϕ_r vary with step size $\Delta\phi \in (0, \pi/2)$, along the elevation direction. While φ_i and φ_r vary with step $\Delta\varphi \in (0, 2\pi)$ along the azimuthal direction. All in all, the total number of rendered objects per BRDF material is $N = 2Sa^2c^2$, with $a = \lceil \frac{\pi}{2\Delta\phi} \rceil + 1$ and $c = \lceil \frac{2\pi}{\Delta\varphi} \rceil$. The set of rendered objects is $\mathcal{B} = \{B_1, \cdots, B_b\}$, with b the number of considered BRDF materials, and each B_i is made of N rendered objects. For the ambient light we used 16 different light probes, see (Debevec, 2008).

Segmentation. Images sample are taken from the ImageNet database (Deng et al., 2009). ImageNet is plenty of objects of several categories, many of which challenging for 3D modeling in terms of concavity, sharpness and specularity. We have sampled some of them, provided they are not occluded. Each testing image is well segmented, choosing manually a main object of interest. We have implemented the level-set based method of (Vese and Chan, 2002), a generalization of the active contours approach considering a multi-level set framework.

5.5 Object properties transfer

In this section we address the following problem. Given examples $\mathbf{X}_B \in \mathbb{R}^{h \times N}$ of image patches of shaded surfaces with varying illumination and curvature, about which we know probe, material, normals, and depth, with $\sqrt{h}$ the size

of the patch, we wish to recover the normals to the surface of a segmented image I_Q, of an unknown object Q, the material it is made of, and the probe. To this end we have to establish a correspondence between the patches of the unknown surface I_Q and the patches of the known r-surfaces $\mathbf{X}_B$, in the synthetic database. We can see the problem under the following perspective. If we consider a hierarchy of properties of a patch, such as surface features like depth, normals, probe, and image features, we can see that each group of features is a scattered realization of a multivariate variable with unknown probability distribution, whose density is an infinite mixture. We thus use a nested Dirichlet process mixture as introduced in (Blei et al., 2010), see also (Rodriguez et al., 2008; Paisley et al., 2015), defining prior distributions on recursive data structures. Assuming that samples of specific patches have been collected for each of J distributions and are contained in vector $\mathbf{y} = (y_1, \ldots, y_J)$, here we consider that each one provides a different distribution modeling mixtures for each group of features, though we deliberately neglect a sharing level. We obtain a k-ary tree of infinite mixtures, such that each level provides classification paths for the specific feature set, within which the next level of features is nested. At each level of the hierarchy each mixture component gathers patches of similar appearance, namely we have Z-patches for depth, $\mathbf{n}$-patches for normals, $\mathbf{p}$-patches for probes and F-patches for visual features.

The idea is that a patch of a segmented image I_Q, showing only image features, is classified according to the highest level of the hierarchy. Then, following the path of the corresponding branch of the tree of infinite mixtures, the probe, the normals and the depth of the patch can be recovered, considering the mean representative of the corresponding component. The advantage of this non-parametric Bayesian approach is that even with 10^4, up to 10^5 patches, it is possible to obtain good classification results. Note that at each node of the tree the infinite mixture estimates parameters, hence components, according to reallocated indices of the parents nodes, ensuring interchangeability at each level, along a path. Note that the number of samples that can be used along a path j at level ℓ is about $N(\prod_{i=1}^{j\ell} n_{c_{j\ell}})^{-1}$, with $n_{c_{j\ell}}$ the number of components in the branch at level ℓ.

A hierarchical model is built for each BRDF in the synthetic database (see Section 5.8 for details). For each model $\mathcal{M}_B$, $B \in \mathcal{B}$, at the base level of the hierarchy the mixture components are generated from the Z-patches, at the next level from the $\mathbf{n}$-patches, then the probes $\mathbf{p}$-patches, and the leaves level is generated from the F-patches. Here the F-patches are obtained by mapping the RGB values into a feature space, so as to extract the features coded in their representation, ensuring statistical independence of the data (Olshausen and Field, 1997; Hinton and Salakhutdinov, 2006). Autoencoders are a popular computational architecture to learn features from data (Bengio et al., 2013; Ngiam et al., 2011), here we introduce a sparse stacked autoencoder, to obtain the F-patches for each BRDF $B \in \mathcal{B}$, which determines the features size from sparsity.

Distribution linking the object image and r-surfaces. Let Y be a multivariate whose density is an infinite Gaussian mixture, with unknown parameters. The nested DPM model we consider is $Y|c_{k,j\ell}, \boldsymbol{\theta}_{k,j\ell} \sim \mathcal{N}(\mu_{c_{k,j\ell}}, \Sigma_{c_{k,j\ell}})$, $k \to \infty$ and $j\ell$ the level on the path j in the tree. Here $c_{k,j\ell}$ indicates the mixture component k, at level ℓ, on the path j and the $\boldsymbol{\theta}_{k,j\ell}$ are in turns independently sampled from an unknown distribution $\boldsymbol{\theta}_{k,j\ell}|G_{j\ell} \sim G_{j\ell}$, on which is placed a Dirichlet process $G_{j\ell} \sim DP(\alpha_\ell G_{0,\ell})$. Here α_ℓ is the concentration parameter, affecting the number of components that will be generated, and $G_{0,\ell}$ is the base distribution, typically the conjugate prior of the observation distribution (for the DPM at each level in a path, we refer the reader to the recent (Blei and Jordan, 2006; Sudderth, 2006) though the models go back to (Ferguson, 1973; Antoniak, 1974)). Assume, now, that the parameters have been computed for each group of features, that a nested DPMs $\mathcal{M}_B$ is obtained for each $B \in \mathcal{B}$, actually each with 4 levels. Each nested DPM has a number of j-paths according to the recursive structure induced by the groups of features. Given a nested DPM for each $B \in \mathcal{B}$ we are concerned with the computation of the data likelihood for a realization $\mathbf{h}_{Q_B}$, of a patch X_Q, whose BRDF has been identified to be B (see below). Once $P(c_{j\ell} = k_{j\ell}|\mathbf{h}_{Q_B}, \mathcal{M}_B)$, is established for the leaf components at level $\ell = 4$, along the path j then, going back along the path and picking the mean value of the nodes in the path, we obtain the most plausible features **p**-patch and **n**-patch matching $\mathbf{h}_{Q_B}$. Note that when the DPM is trained, the realizations of Y are the patch features $\mathbf{h}_B$ of the X_B in the synthetic database. To compute the nested DPM we have used conjugate priors and an extension of (Jain and Neal, 2004), see also (Sudderth, 2006; Natola et al., 2015b).

Stacked sparse autoencoder for each BRDF. Let $\Omega \subseteq \mathbb{R}^h$ be the data space, H the feature space, and $X \in \Omega$ be a patch. Autoencoders (Munro and Zipser, 1989; Ngiam et al., 2011) provide a structured representation of the sample data, by estimating an encoding map $f : \Lambda \times \Omega \mapsto H$, and a decoding map $g : H \times \Lambda \mapsto \Omega$. Features generated by an autoencoder $\beta(B)$ take values $\mathbf{h} = f(\Lambda_\beta, X) = \sigma(W_{in}X + \mathbf{b}_{in})$. Optimization for minimizing the loss function is here obtained by the orthant projection method (Andrew and Gao, 2007; Schmidt et al., 2012). The result of the optimization for the stacked autoencoder are the parameters $\Lambda_\beta^{(1)} \cup \Lambda_\beta^{(2)}$.

The final features for patches $\mathbf{X}_B$, for $B \in \mathcal{B}$, is $\mathbf{h}_B = \sigma(W_{in}^{(2)}\mathbf{h}_B^{(1)} + \mathbf{b}_1^{(2)} \otimes \mathbf{1}_{1\times M})$, of size $k \times M$; here $\mathbf{h}_B^{(1)} = \sigma(W_{in}^{(1)}\mathbf{X}_B + \mathbf{b}_1^{(1)} \otimes \mathbf{1}_{1\times M})$ are the lighter feature values, and $\otimes$ is the Kronecker product.

On the other hand, let $\mathbf{X}_Q = (X_{Q_1}, \ldots, X_{Q_K}) \in \mathbb{R}^{h\times K}$ be the K patches of I_Q (segmented image of Q). The feature set for I_Q is:

$$
\begin{aligned}
&H_{Q/B} = \\
&\{\mathbf{h}_Q = \sigma(W_{in}^{(2)}\sigma(W_{in}^{(1)}\mathbf{X}_Q + \mathbf{b}_1^{(1)} \otimes \mathbf{1}_{1\times K}) + \mathbf{b}_1^{(2)} \otimes \mathbf{1}_{1\times K})| \\
&\quad (W_{in}^{(2)}, W_{in}^{(1)}, \mathbf{b}_1^{(2)}, \mathbf{b}_1^{(1)}) \in \Lambda_\beta^{(1)} \cup \Lambda_\beta^{(2)}, \forall B \in \mathcal{B}\}
\end{aligned}
\tag{5.1}
$$

These features are obtained by evaluating each stacked autoencoder $\beta(B), B{\in}\mathcal{B}$, at $\mathbf{X}_Q$. To choose one, consider the average features for $B{\in}\mathcal{B}$: $\mathbf{s}{=}1/M\sum_{\forall X_B}\mathbf{h}_B$. Let $\varepsilon(x){=}-\log(x)$, be the Burg entropy, then according to (Csiszár, 1996) we obtain Bregman divergence to measure similarity between the object features and $\mathbf{s}$:

$$\begin{aligned}&\mathbf{X}_Q \in B^\star \text{ if } B^\star = \arg\min_B d(\mathbf{X}_Q, B), \quad \text{with}\\ &d(\mathbf{X}_Q, B){=}\sum_{\forall \mathbf{h}_Q \in H_{Q/B}} (\varepsilon(\mathbf{s}){-}\varepsilon(\mathbf{h}_Q)){-}\nabla\varepsilon(\mathbf{h}_Q)(\mathbf{s}{-}\mathbf{h}_Q)\end{aligned} \tag{5.2}$$

This results in a full identification of the specific BRDF B for each X_Q, as the material of the patch. Once the BRDF B is chosen, the features $\mathbf{h}_Q$ are the specific realizations of the multivariate Y. Hence the nested DPM can be applied, as gathered in the previous paragraph, in order to obtain the sought for properties to be transferred to X_Q.

5.6 Bas-relief modeling of objects

In this section we present the method for modeling an object shape, given the information obtained from the inference, described in Section 5.5. Accordingly, we are given a number of patches $\mathbf{X}_Q$ covering the segmented image of object Q, the normal field transferred from some X_B, and the position of the top left corner within the domain Ω. Note that the patches are not overlapping.

Object modeling using normals and curvatures Here we define a binary mask $A{\subset}\mathbb{R}^2$ for image I_Q by the mapping $\nu{:}\Omega{\mapsto}\{0,1\}$. The surface, parametrized by the function $\mathbf{w}{:}A{\mapsto}\mathbb{R}^3$, where $\mathbf{w}(u,v)$ is the vector $[x(u,v), y(u,v), z(u,v)]^\top$, is obtained by minimizing an energy functional $\mathcal{G}(\mathbf{w})$. The energy functional $\mathcal{G}(\mathbf{w})$ is defined by the first and second fundamental forms (Terzopoulos et al., 1987), and it embeds surface stretching and bending, plus external forces F acting on it (Ntouskos et al., 2015b).

To correctly identify the external forces we compute the mean curvature $\kappa(u,v)$ for each $(u,v){\in}A$, given the normal $\mathbf{n}(u,v)$ at each point of the surface, as estimated by the N-DPM, see Section 5.5. The external forces are needed to sculpt the surface inflation and are of the form $F(u,v) = \text{sign}(\kappa(u,v))q\mathbf{n}(u,v)$, with $q{\in}\mathbb{R}^+$. The scheme for finding the solution $\mathbf{w}(\cdot)$ is based on the Finite Element method, as described in (Strang and Fix, 1973), applied to the Euler-Lagrange equations associated to the functional $\mathcal{G}(\mathbf{w})$. Furthermore, we require that each normal to the surface $\mathbf{w}(u,v)$ is a unit vector along $\mathbf{w}_u{\times}\mathbf{w}_v$, with $\mathbf{w}_u, \mathbf{w}_v$ the partial derivatives of $\mathbf{w}$. These conditions are imposed as follows:

$$\begin{aligned}\mathbf{n}(u,v)\cdot\mathbf{w}_u(u,v) &= 0\\ \mathbf{n}(u,v)\cdot\mathbf{w}_v(u,v) &= 0.\end{aligned} \tag{5.3}$$

To linearize the constraints in the model parameters, we add to $\mathbf{w}$ further degrees of freedom including partial derivatives: $\hat{\mathbf{w}}(u,v) = [x, y, z, x_u, y_u, z_u, x_v, y_v, z_v]^\top$.

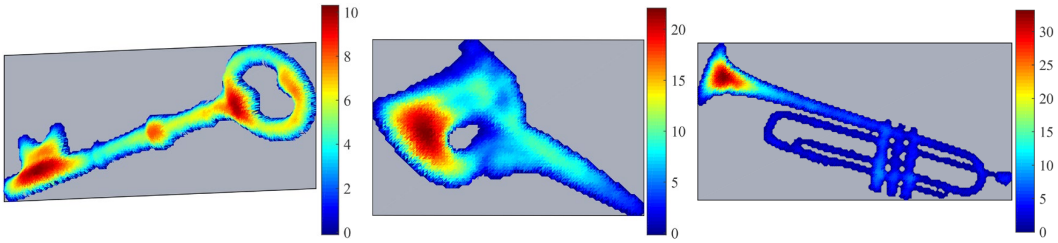


Figure 5.3. Modeled surfaces from the segmented images of a key, a mask and a trumpet.

The constraints for (u, v), (5.3), can now be formulated as follows:

$$\begin{bmatrix} 0 & 0 & 0 & \mathbf{n}^x & \mathbf{n}^y & \mathbf{n}^z & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{n}^x & \mathbf{n}^y & \mathbf{n}^z \end{bmatrix} \hat{\mathbf{w}}(u,v) = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

with $\mathbf{n}^x, \mathbf{n}^y, \mathbf{n}^z$ the components of $\mathbf{n}(u, v)$ in the x, y, z directions. The constraints in linear form can be expressed as a matrix equation $D\mathbf{U}{=}C$, with $D{\in}\mathbb{R}^{2\omega\times l}$, $C{\in}\mathbb{R}^{2\omega\times 1}$, and $\mathbf{U} = [\hat{\mathbf{w}}(u_1, v_1)^\top, \ldots, \hat{\mathbf{w}}(u_\omega, v_\omega)^\top]^\top{\in}\mathbb{R}^{l\times 1}$ the vector including the total number l of d.o.f. of the system, and ω being the total number of points inside A. The quadratic minimization problem becomes:

$$\min_{\mathbf{U}} \left\{ \mathbf{U}^\top K\mathbf{U} - F^\top \mathbf{U} + (D\mathbf{U}-C)^\top \Gamma (D\mathbf{U}-C) \right\}, \tag{5.4}$$

with $K{\in}\mathbb{R}^{l\times l}$ the stiffness matrix, (Strang and Fix, 1973), $F{\in}\mathbb{R}^{l\times 1}$ the vector of the external forces and $\Gamma{\in}\mathbb{R}^{2\omega\times 2\omega}$ a diagonal matrix with elements the weight $\gamma_i{\in}\mathbb{R}$ of each constraint, for $i{=}1, \ldots, \omega$, defined as $\Gamma = \text{diag}(\gamma_1, \gamma_1, ..., \gamma_N, \gamma_N)$. To constrain the solution at the boundary ∂A, homogeneous Dirichlet conditions are applied to the PDE problem. Once the solution $\mathbf{U}$ is computed, the surface and corresponding mesh, obtained from the triangulation over A, are reconstructed. Some modeled surfaces are shown in Figure 5.3.

5.7 Photo-consistency and smoothness

To resolve irregularities of the surface due to noise and outliers we refine the initial surface. Function $z(u, v)$ provides the height of the initial surface, as discussed in Section 5.6. We model the image $\hat{I}(z)$ considering the surface $z(u, v)$ rendered with the recovered probe and BRDF. The goal of the surface refinement is to enforce photo-consistency with the given image while smoothing out the initial surface. The photo-consistency error between the modeled image $\hat{I}$ and the shading of the surface I_s in the given image is given by

$$E_{photo}(z) = \|I_s - \hat{I}(z)\|_1. \tag{5.5}$$

As we consider objects of specular BRDF, intensity values of the images are strongly affected by the surrounding environment. We considered the reflected

environment as a texture modulating the intensities of the imaged object and we approximate the shading image I_s by separating the shading and specularity components of the object via Retinex (Land and McCann, 1971).

Smoothing of the initial surface is achieved by applying total generalized variation (TGV) regularization of the height map $z(u, v)$ corresponding to the initial surface. TGV regularization encourages a piece-wise smooth reconstruction of the height map with polynomial terms up to order η (Bredies et al., 2010; Burger and Osher, 2013). This leads to

$$E_{depth}(z) = TGV^{\eta}(z). \tag{5.6}$$

Finally, to avoid excessive distortion of the surface, due to the presence of outliers in the shading image I_s, we require that the normals of the refined surface are similar to the ones of the initial surface. Letting $\mathbf{n}(u, v)$ be the normal of the surface at the point (u, v) and $\mathbf{n}_0(u, v)$ the initial normal at the same point, we consider the following fidelity term

$$E_{norm}(\mathbf{n}) = \|\mathbf{n}(u, v) - \mathbf{n}_0(u, v)\|_1. \tag{5.7}$$

The final surface is obtained by minimizing the resulting energy-like functional, for TGV^0 this is:

$$E(z)=E_{depth}(z) + w_1 E_{photo}(\hat{I}(z)) + w_2 E_{norm}(\mathbf{n}(z)), \tag{5.8}$$

with w_k the weights of the fidelity terms, $k = 1, 2$.

The function (5.8) is non-convex due to the terms E_{photo} and E_{norm}. We relax the problem by considering a local linear approximation of the $\mathcal{S}^2$ manifold as described in (Zeisl et al., 2014). Let $\mathbf{n}_l$ be the linearization point of the normal field, and $T = \text{null}(\mathbf{n}_l)$, then $\mathbf{n}(z) = T\nabla z + \mathbf{n}_l$, up to a normalizing constant. Integrability of the normal field (Papadhimitri and Favaro, 2013; Reddy et al., 2009) is automatically satisfied in this case. The functional of the relaxed problem is:

$$\begin{aligned} E(z, \zeta) = \int_{\Omega} & |\nabla z| + w_1 \|T\nabla z + \mathbf{n}_l - \mathbf{n}_0\| \\ & + \frac{1}{2\theta}(\zeta - z)^2 + w_2 |I_s - \hat{I}(\zeta)| du dv. \end{aligned} \tag{5.9}$$

The auxiliary variable ζ is purposefully added in (5.9) to separate the photo-consistency from the rest of the terms, in so separating the problem into two distinct minimization sub-problems. At each iteration the minimizer of the photo-consistency term is estimated by point-wise search, while a minimizer with respect to z is identified by primal-dual optimization (Chambolle and Pock, 2010).

Considering the part of (5.9) depending only on z, we obtain its primal-dual form by applying the Legendre-Fenchel transformation. Let $\mathcal{P}$ be the convex

set obtained from the union of L_1 balls, D the discretized gradient operator, and $\mathbf{z}, \boldsymbol{\zeta}, \bar{\mathbf{n}}$ the vectorized variables corresponding to $z, \zeta, \mathbf{n}$ respectively, then the primal-dual form of (5.9) is:

$$\max_{\mathbf{p},\mathbf{q}\in\mathcal{P}} \quad \frac{1}{2\theta}\|\boldsymbol{\zeta}^* \;-\; \mathbf{z}\|^2 \;+\; \langle \mathbf{p}, \mathrm{D}\,\mathbf{z}\rangle \;+\; w_1\langle \mathbf{q}, \mathrm{T}\,\mathrm{D}\,\mathbf{z} \;+\; \bar{\mathbf{n}}_l \;-\; \bar{\mathbf{n}}_0\rangle. \quad (5.10)$$

Choosing suitable step sizes $\sigma, \tau > 0$, a saddle point is found by the proximal point iterations summarized below:

$$\begin{aligned}
\mathbf{p}^{(k+1)} &= \Pi_{\mathcal{P}}\left(\mathbf{p}^{(k)}+\tau\,\mathrm{D}\,\hat{\mathbf{z}}^{(k)}\right),\\
\mathbf{q}^{(k+1)} &= \Pi_{\mathcal{P}}\left(\mathbf{q}^{(k)}+\tau w_1(\mathrm{T}^{(k)}\,\mathrm{D}\,\hat{\mathbf{z}}^{(k)}+\bar{\mathbf{n}}_l^{(k)}-\bar{\mathbf{n}}_0)\right),\\
\mathbf{z}^{(k+1)} &= (1+\tfrac{\sigma}{\theta^{(k)}})^{-1}\left(\mathbf{z}^{(k)}+\tfrac{\sigma}{\theta^{(k)}}\boldsymbol{\zeta}^*\right.\\
&\quad \left.-\,\sigma\,\mathrm{D}^\top(\mathbf{p}^{(k+1)}+w_1{\mathrm{T}^{(k)}}^\top\mathbf{q}^{(k+1)})\right),\\
\hat{\mathbf{z}}^{(k+1)} &= 2\mathbf{z}^{(k+1)}-\mathbf{z}^{(k)},\\
\bar{\mathbf{n}}_l^{(k+1)} &= \Pi_{S^2}(\mathrm{T}^{(k)}\,\mathrm{D}\,\mathbf{z}^{(k+1)}+\bar{\mathbf{n}}_l^{(k)}),
\end{aligned}$$

with $\mathrm{T}^{(k)}$ a matrix formed by the the null spaces of the corresponding vectors $\bar{\mathbf{n}}_l^{(k)}$, Π_X the projection on set X, and w_k as mentioned in (5.8). θ decreases at each iteration, enforcing the variables $\boldsymbol{\zeta}$ and $\mathbf{z}$ to converge, approximating in this way a solution of the original minimization problem.

The refinement produces smooth surfaces while preserving sharp discontinuities of the initial surface supported by the appearance of the object in the image.

5.8 Experiments and results

Unsupervised learning experiments. We consider the following BRDFs: aluminum, brass, PVC, steel and plastic. For each material up to N=430 r-surfaces are generated, and about 23.30×10^4 patches obtained. Transformation of patches into feature space lasts 32.12×10 sec., for each $\beta(B)$. DPM training lasts about 60.40×10^4 sec. for each B. These on a computer equipped with four Xeon E5-2643 3.7GHz CPUs and 64GB RAM.

MSE prediction error for autoencoders is shown in Figure 5.4. Material choice (eq. 5.2) is 100% correct. To evaluate the accuracy of components prediction for the observed object with the DPM, we use 3D models with computed normals and rendered with BRDF (Figure 5.6). Results are given in Figure 5.5, where the size N of the r-surfaces samples, varies from 48 to 430. Mixtures components range from a minimum of 18×10 to a maximum of 27×10^2. Ground truth (GT) objects are also used to evaluate the NMSE of mean normals between each X_Q and each representative X_B of the chosen DPM component, Figure 5.5 right.

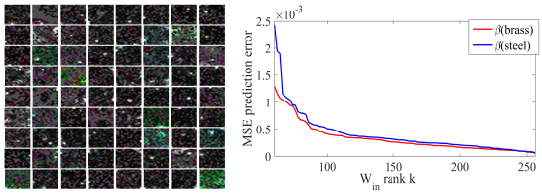


Figure 5.4. On the left the deep features predicted by $\beta(brass)$, with rank k=72, m=256. On the right autoencoders $\beta(steel)$ and $\beta(brass)$ MSE prediction error, according to reduced $W_{in}^{(2)}$ rank. Rank k is varied from a 22.6% reduction, up to no reduction.

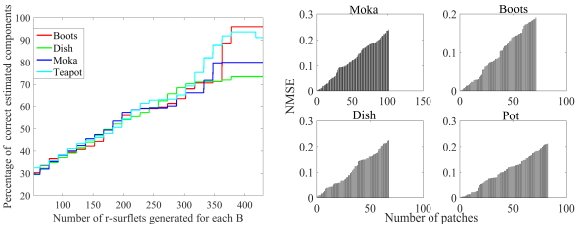


Figure 5.5. On the left components prediction accuracy for the ground truth objects shown in Figure 5.6, varying the size of the sampled r-surfaces. On the right accuracy w.r.t. mean normals.

Synthetic data We examine first the performance of the framework using synthetic images for which the ground truth is available. We render various 3D models using the BRDFs of the materials we consider in this paper, taken from the MERL dataset (Matusik et al., 2003). Renderings using the measured BRDFs are obtained by using a data-driven light closure of the Cycles 3D render engine in Blender. Photorealistic views of the 3D models are composed by using suitable HDR light probe images for simulating surrounding environments. Moreover, we compute the ground truth depth map and the normal map of the rendered object with respect to the current view, by using specialized OSL shaders.

We apply our method on these synthetic views and compare the results with the ground truth. For evaluating the error in the depth field we use the Z-MAE measure (Barron and Malik, 2015), normalized with respect to the object bounding box diagonal. For the error of the normal field we use the median angular error (N-MAE) (Barron and Malik, 2015), and the mean-squared error of the normal field (N-MSE). The shading error is evaluated using the L-MSE error introduced in (Grosse et al., 2009), considering a window of size 20. Finally, the error between the modeled surface and the GT object is measured using the normalized Hausdorff distance (Aspert et al., 2002). The average values are computed by taking the geometric mean of the values, as in (Barron and Malik, 2015). The results are shown in Table 5.1, and images of the rendered 3D objects and the surfaces obtained from our method are

presented in Figure 5.6. In the same figure, the absolute shading distance and the distance between the meshes are also visualized. The images are best viewed in color and on screen.

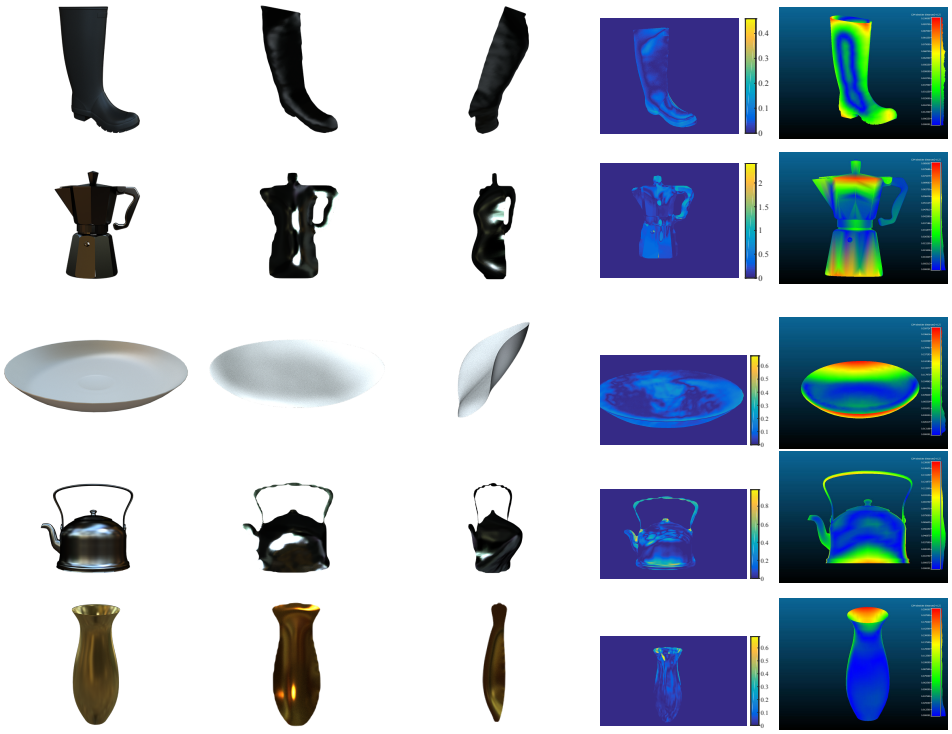


Figure 5.6. Models with ground truth. **1st col.** GT 3D model with BRDF; **2nd col.** modeled surface with BRDF; **3rd col.** rotated view; **4th col.** shading difference; **5th col.** Hausdorff distance.

Object	Z-MAE	N-MAE	N-MSE	L-MSE	Hausdorff	Average
boot	0.0749	0.6397	0.4052	0.0012	0.0460	0.1160
moka pot	0.0632	0.4260	0.2842	0.0808	0.0340	0.0640
dish	0.2434	0.3060	0.2426	0.0009	0.0594	0.0627
teapot	0.1265	0.4325	0.3976	0.0348	0.0713	0.1401
vase	0.0494	0.1737	0.1990	0.0193	0.0721	0.0750
Average	0.0936	0.3626	0.2944	0.0090	0.0544	0.0867

Table 5.1. Synthetic images results.

The results show that our algorithm produces plausible surfaces of the imaged object from a single image. The material was successfully recognized every time, while the average value of the median angular error is about 22°. We observe that the shading distance does not always follow the angular and depth error, justifying the use of different error metrics for assessing the modeled surface quality. Three of the objects have significant concave parts (boot, plate, vase) which are evident also in the modeled surfaces. Finally, we

see that the metallic objects although showing an increased shading error, due to residual reflections of the environment, are still modeled faithfully, according to the shape metrics.

MIT dataset For an evaluation of our method with respect to publicly available data we use the MIT intrinsic image dataset (Grosse et al., 2009), as augmented in (Barron and Malik, 2015) to include the shape of each object. We consider the objects *apple, potato, teabag1, teabag2, paper1* as they exhibit specularity and/or concavities. The objects of this dataset are made of different materials with respect to the ones existing in the MERL BRDF dataset. To overcome this problem we combine the shading and specularity images of the objects to obtain new composite images without texture. The algorithm recognizes *plastic* as the most similar material to the shaded-only object. Figure 5.7 shows the reference images and the modeled surfaces for each object of the dataset.

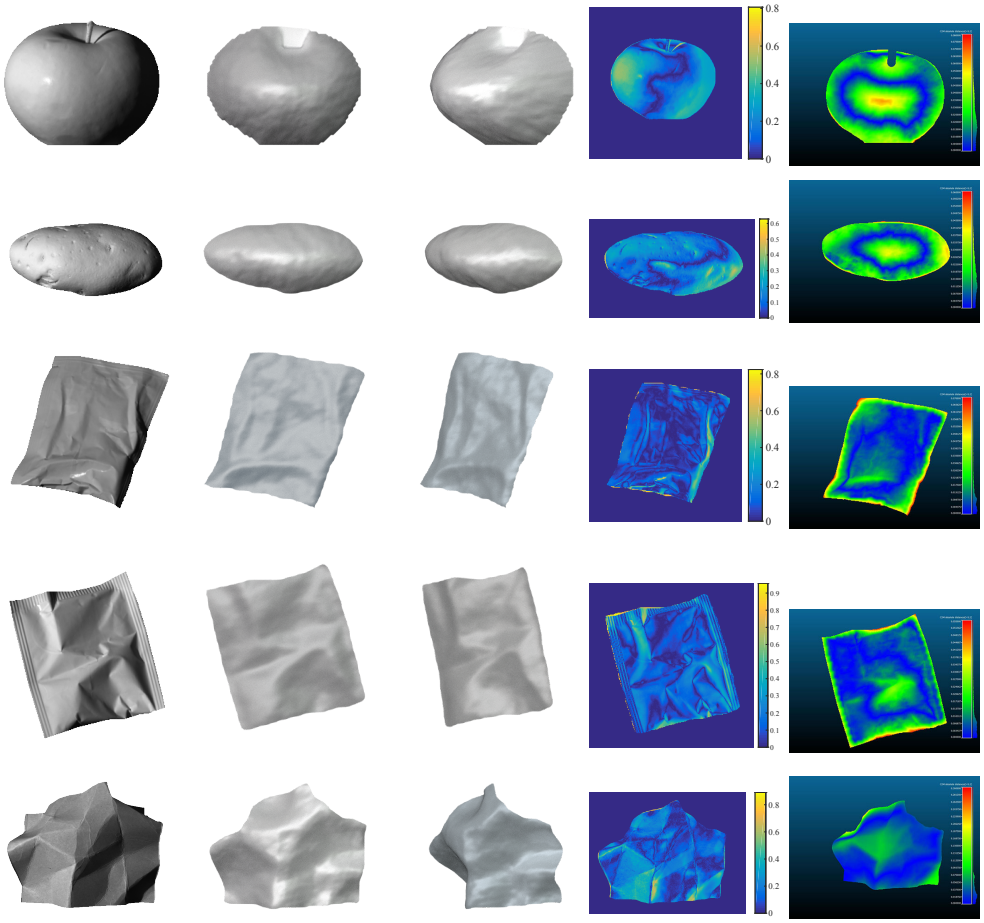


Figure 5.7. MIT dataset. **1st col.** reference image; **2nd col.** modeled surface with BRDF; **3rd col.** rotated view; **4th col.** shading distance (L-MSE); **5th col.** Hausdorff distance.

Algorithm	Z-MAE	N-MAE	S-MSE	L-MSE	Avg.
Ours	**7.0197**	**0.2692**	0.0261	0.0174	0.1712
Ours no FC no S	26.9816	0.5872	0.0394	0.0217	0.3412
Ours only contour (SfC)	37.1768	0.7728	-	-	-
Retinex+SIFS(Barron and Malik, 2015)	17.1914	0.9361	0.0006	0.0019	0.0654
SIFS(Barron and Malik, 2015) (grey, lab. light)	20.1445	0.9772	**0.0005**	**0.0017**	**0.0640**

Table 5.2. Results of full and ablated model on MIT dataset (Grosse et al., 2009).

Table 5.2 compares our results with (Barron and Malik, 2015). As the input images are albedo-less, SIFS (Barron and Malik, 2015) was used as a baseline. For the comparison the Z-MAE metric is reported with no normalization and the S-MSE metric (Barron and Malik, 2015) is also considered. On one hand the results show that SIFS achieves better results on shading metrics. This is reasonable, since (Barron and Malik, 2015) directly optimizes over the rendering error, while in our approach photo-consistency is sought after shape has been recovered. Still, our method achieves higher accuracy on shape metrics, since it primarily recovers the surface normals. On comparing the shape recovered with the two approaches one can notice that (Barron and Malik, 2015), due to the Lambertian assumption, distorts shape near reflections and specularities, trying to interpret intensity changes as changes in shape. Additionaly, (Barron and Malik, 2015) cannot always capture concavity of the surface (e.g. the bowl of the spoon in Figure 5.8). Note that in Table 5.2 we considered also a pre-processing with Retinex before applying SIFS, which helps in reducing specularities, leading to better results in terms of shape, slightly penalizing the shading distance. Table 5.2 presents also ablated versions of our method, highlighting the importance of surface refinement.

Modeling of ImageNet objects We have manually selected from the ImageNet dataset (Deng et al., 2009) images of objects made from the materials described above. The 3D surfaces of the visible parts of these objects are computed with the proposed framework. Figure 5.8 shows the selected images together with renderings of the recovered surface as well as the computed depth and normal maps before and after refinement. Comparison with the results of (Barron and Malik, 2015) is also provided. We observe that the modeled surfaces closely resemble the reference objects, when viewed from the image vantage point with the recognized probe and BRDF. This is also evident by the values of the shading difference and the L-MSE metric, reported in Table 5.3.

Algorithm	Concave spoon	Glove	Trumpet	Key	Funnel	Convex spoon	mask	Average
Ours	0.0792	0.0559	0.0571	0.0271	0.0189	0.0321	0.471	0.0570
(Barron and Malik, 2015) (color, natural ill.)	0.0669	0.0097	0.1600	0.0204	0.0072	0.0337	0.0077	0.0169

Table 5.3. L-MSE for ImageNet objects.

5.9 Conclusions

We proposed a novel approach for BRDF aware modeling of 3D objects from a single image. The contributions of the paper are twofold. On the one side, we are able to fully model non-Lamberitan surfaces with either concave or sharp parts, with limited error both in shading and shape. On the other side, we have proved that the normal field of the surfaces to be modeled can be learned from renderings of different objects surfaces. The contribution builds on three main achievements. The first, is that we can represent the material reflectance and specular properties, basing on deep features, as a hierarchy of features that can be transferred via a nested Dirichlet process mixture to an unknown surface. The second, is that the normal field can be used to define an external force needed to sculpt a deformed surface into a refined shape representation of the unknown object. Finally, we contribute with a new method based on TGV to enforce photo-consistency between the generated surface and the appearance of the object in the image. These results prove to be very promising, despite the whole process seems to be still complex and time demanding.

In future work we will examine the steps needed to retrieve the geometry of the full object, even if a prior is needed. Moreover, we will extend the categories our model can afford and simplify the whole framework.

Acknowledgments

Supported by the EU FP7 TRADR (609763) and the EU H2020 SecondHands (643950) projects. We thank the anonymous reviewers for their insightful comments.

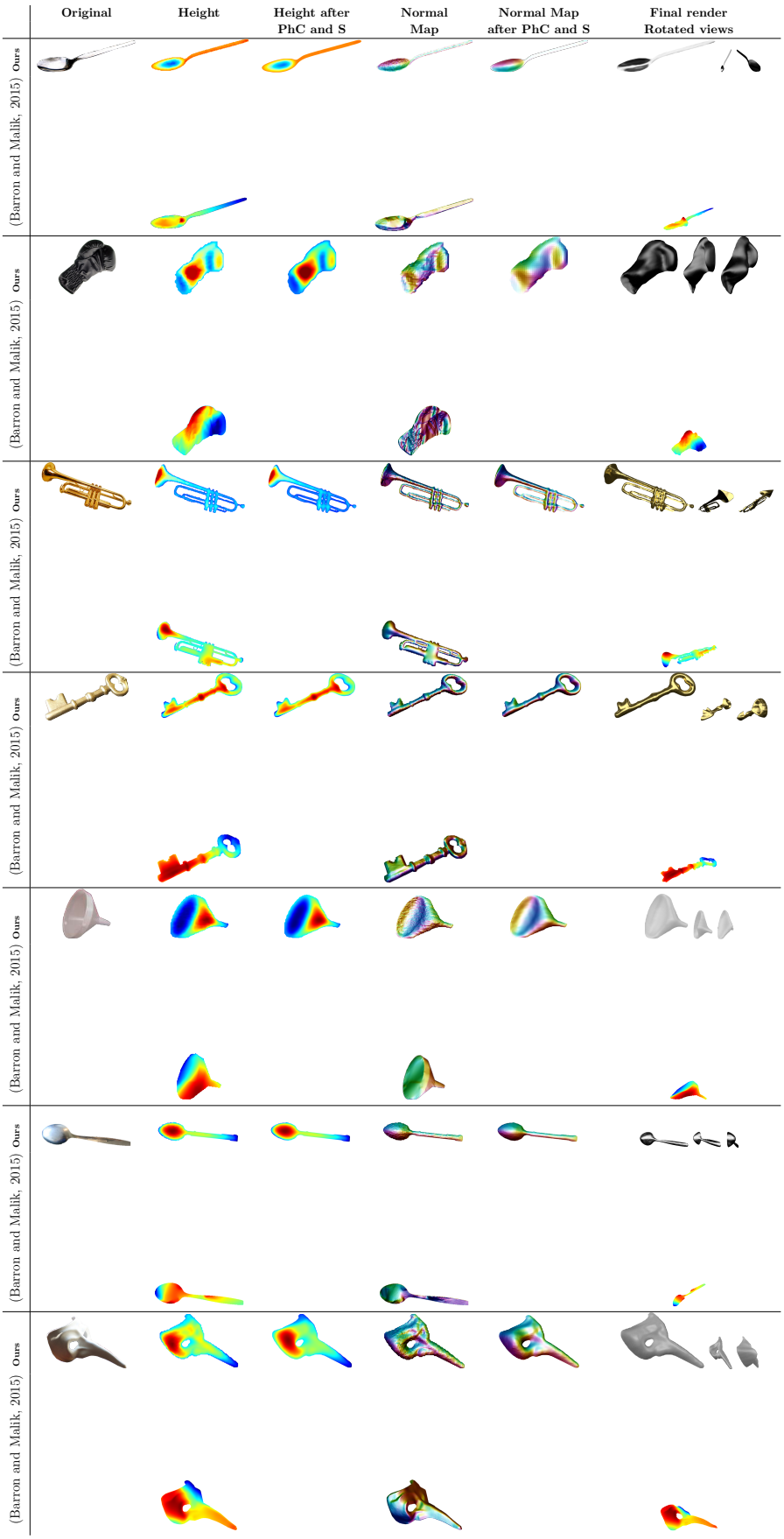


Figure 5.8. Visual comparison between height and normal maps estimated before and after the photo-consistency (PhC) and smoothing (S). Visual comparison with (Barron and Malik, 2015) for the height and normal maps.

Chapter 6

Bayesian Image based 3D Pose Estimation

Marta Sanzari[1], Valsamis Ntouskos[1], & Fiora Pirri[1]

[1] ALCOR LAB, Dipertimento di Ingegneria Informatica Automatica e Gestionale, Sapienza University of Rome

Published: In European conference on computer vision (pp. 566-582). Springer, Cham, 2016.

Statement of Contributions of Joint Authorship

Marta Sanzari (Candidate):
Writing and compilation of manuscript, established methodology, data analysis, preparation of tables and figures.
Valsamis Ntouskos (Research Colleague):
Writing and compilation of manuscript, established methodology, data analysis, preparation of tables and figures, co-author of manuscript.

Fiora Pirri (Principal Supervisor):
Supervised and assisted with manuscript compilation, editing and co-author of manuscript.

Here we present the work done for 3D human pose estimation, published at the European Conference on Computer Vision 2016, which introduces a method for 3D human pose estimation from a single image based on a hierarchical Bayesian non-parametric model, decomposing the human skeleton in groups. A standard human activity dataset is employed, containing both information about 2D and 3D skeleton joints, and a hierarchical model connecting 3D poses and 2D visual features (namely PHOG features) was built. In order to

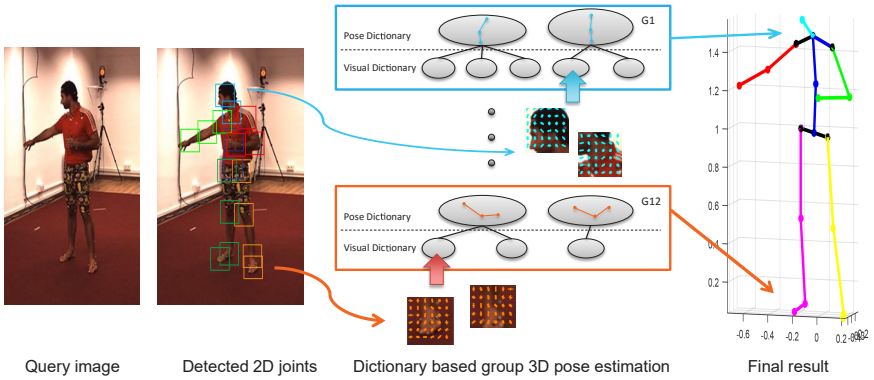


Figure 6.1. Method overview; 3D pose estimation given a query image.

infer the 3D skeleton data from new unseen images, a state-of-the-art method for 2D skeleton joints extraction from images was employed. From image regions sourrounding the 2D joints, PHOG features are extracted, and 3D joints are inferred from the learned pose dictionary of each skeleton group. The final complete 3D human pose, a full-pose consistency condition have been formulated.

This Chapter is an exact copy of the conference paper referred to above.

6.1 Abstract

We introduce a 3D human pose estimation method based on hierarchical Bayesian non-parametric models. Considering a decomposition of the human skeleton joints into groups, our model generates a dictionary representative of the motion and the appearance of each group. Given a query image, the learned dictionary is used to estimate the likelihood of the group pose based on its visual features. The pose of the full-body is reconstructed taking into account the pose consistency of the connected groups. The results show that the proposed approach is able to accurately reconstruct the 3D pose of previously unseen subjects.

6.2 Introduction

Human pose estimation from images has been considered since the early days of computer vision and many approaches have been proposed to face this quite challenging problem. A large part of the literature has concentrated

on identifying a 2D description of the pose mainly by trying to estimate the positions of the human joints in the images. Recently, attention has been shifted to the problem of recovering the full 3D pose of a subject either from a single frame or from a video sequence. Despite this is an ill-posed problem due to the ambiguities emerging by the projection operation, the constraints induced by both human motion kinematics and dynamics have facilitated the recovery of some accurate 3D human pose estimation.

In this work we approach the problem of 3D pose estimation from a single image building a hierarchical framework based on Bayesian non-parametric estimation. A schema of the framework is shown in (Fig. 6.3). Following the schema flow, we divide the human body into different parts and we study the idiosyncratic motion behavior of each part independently from the others. In this way we learn the principal motion modes of each part. Each body part is specified by a group of joints, and its motion is represented by pose features obtained by the principal motion direction on the $SE(3)$ manifold with respect to a reference pose. As a natural reference pose we consider the "Vitruvian man" pose presented in Fig. 6.2 together with the selected groups.

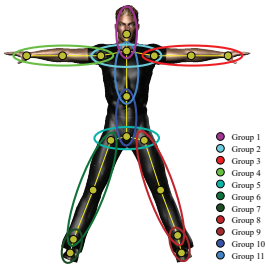


Figure 6.2. "Vitruvian" pose with defined groups.

The visual features for each group are the PHOG features of (Bosch et al., 2007), which are computed using the state-of-the-art approach of (Yang and Ramanan, 2013). Assuming a correspondence between the visual and pose features both the space of visual features and pose features are partitioned, in such a way that from the visual features it is possible to accede to the non observed pose features. These nested partitions are built up for each group with a hierarchical non-parametric Bayesian model, designed purposefully to deal with the inverse projection problem, from 2D to 3D. Indeed, the goal is to recover the unknown human poses just from the available visual features, since visual features are the only available observations.

The hierarchical model is based on two nested countably infinite mixtures of normal distributions. The first level builds a dictionary of 3D human poses by considering various examples of 3D human poses taken from a large number of motion sequences, while the second level takes into account the corresponding images obtained from a number of view points. Indeed, the dictionary is built by partitioning the space of 3D poses with a Dirichlet process mixture model (DPM). The partition is defined on the space of poses specified by the principal motion directions on the $SE(3)$ manifold. The nested part of the model builds the visual dictionary on top of the pose dictionary, and it is also based on Dirichlet process mixture models. Here the mixture processes the PHOG (Bosch et al., 2007) features extracted from a window centered at the 2D position of each joint in the given image.

Based on the learned dictionary 3D pose estimation is performed as follows

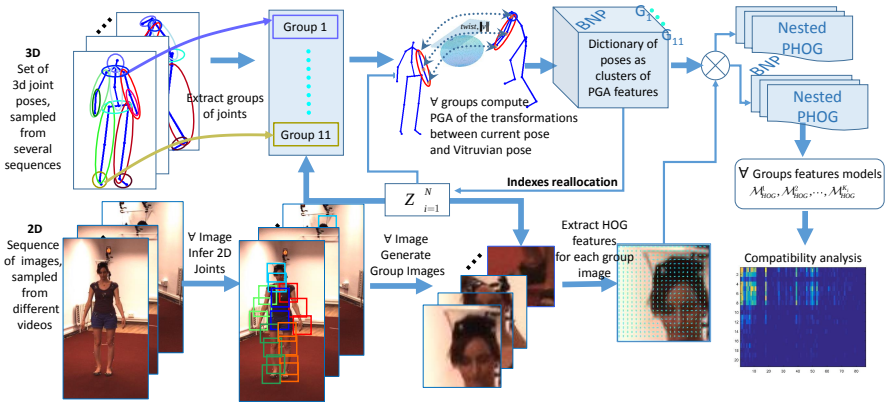


Figure 6.3. Schematic representation of the proposed hierarchical model.

(Fig. 6.1). Given a query image we extract the 2D positions of the joint in the image using a state-of-the-art approach (Yang and Ramanan, 2013) and compute the corresponding PHOG features for each group. From these features we infer the most likely cluster of the visual dictionary, which in turns indicates the cluster of 3D poses with the highest probability for the given group. The final 3D pose is reconstructed by assembling together the most representative poses of the selected clusters for each group. Clusters are selected considering also the compatibility between the group poses.

In the following, Section 6.3 discusses related work and Section 6.4 the structure of the training and testing data, and preliminaries. Section 6.5 presents the architecture of the proposed model and how pose estimation is performed. In Section 6.6 we present the results obtained with our method in comparison with state-of-the-art 3D pose estimation approaches. Finally, Section 6.7 discusses conclusions and future work.

6.3 Related Work

Human pose estimation (HPE) has been extensively studied during the years by considering videos, 2D images and depth data, (Liu et al., 2015; Hen and Paramesran, 2009; Poppe, 2007). There exist several open problems; among them we mention variations in human appearance, clothing and background, arbitrary camera view-point, self-occlusions and obstructed visibility, ambiguities and inconsistency in the estimated poses.

Different features can be chosen to describe the different types of data. Focusing on 2D input data, some works assume the 2D body joints locations already given (Akhter and Black, 2015a), while others extract features from silhouettes such as HOG (Dalal and Triggs, 2005), PHOG (Bosch et al., 2007), SIFT (Lowe, 1999) and shape context (Belongie et al., 2002), or dense trajectories (Zhou and De la Torre, 2014).

In detail, concerning 3D HPE from videos, very recently (Zhou and De la Torre, 2014) introduced a spatio-temporal matching (STM) among 3D Motion Capture (MoCap) data and 2D feature trajectories providing the estimated camera view-point and a selected subset of tracked trajectories. In our approach, instead, as in (Simo-Serra et al., 2012; Wang et al., 2014a), body parts in 2D are detected by using the algorithm introduced in (Yang and Ramanan, 2013).

In the last years many works have approached the estimation of the poses via deep learning as in (Li and Chan, 2014; Tompson et al., 2014; Ouyang et al., 2014; Toshev and Szegedy, 2014). In Zhou et al. (Zhou et al., 2016) a convolutional neural network is used to estimate the 2D joint locations in the image. 3D pose sequences are then estimated via an EM algorithm over the entire video by considering a sparse model of 3D human pose in input where each 3D body pose is represented by a linear combination of a predefined basis of poses. Wang et al. (Wang et al., 2014a) propose an overcomplete dictionary of poses learned from 3D human poses and HPE is managed by minimizing an L_1 norm error between the projection of the 3D pose and the corresponding 2D detection, optimizing via alternating direction method. In (Sigal and Black, 2006), body part detectors provide proposals for the location of 2D pose of visible limbs. The 2D pose is then refined via non-parametric belief propagation and the corresponding 3D pose is estimated by learning the parameters of a mixture of experts model.

In (Agarwal and Triggs, 2006) a relevance vector machine is proposed to learn a reconstruction function that is a linear combination over a set of basis functions. The authors extract shape descriptors from a set of 2D images and the corresponding 3D poses. (Mori and Malik, 2006) store a set of different images and full body poses, both in 2D, together with the corresponding viewpoint. A test image is directly matched with all the training images via the shape context matching procedure. The 3D positions are then estimated via the Taylor's approach (Taylor, 2000). Differently from ours, their methods is instance-based, which is not feasible for a real-time application, without also the possibility of generalizing over the training images.

Assuming that joint positions are already given in 2D with the corresponding image, (Akhter and Black, 2015a) propose to learn pose-dependent joint angle limits from a MoCap dataset, to form a prior for estimating the 3D poses, together with the camera parameters. A tracking-by-detection technique is used in (Andriluka et al., 2010) to collect a small number of consecutive video frames. A novel class of descriptors, called tracklets, is defined and 3D poses are recovered from them. In (Lehrmann et al., 2013), human pose is estimated via a non-parametric Bayesian network and structure learning, considering the dependencies of body parts. In our approach, instead, nested non-parametric clustering is considered to find relations among the appearance and the 3D pose of each body part. As in (Lehrmann et al., 2013), our approach is able to generalize over the observed data so as to generate new poses never seen before.

In (Ionescu et al., 2014), besides the construction of a large dataset, a benchmark among various HPE approaches is performed. (Pons-Moll et al., 2014) use boolean relationships between body components, called posebits, for training an SVM for retrieving the 3D body pose. Finally, (Yasin et al., 2015) consider annotated 2D images and MoCap data as independent input data to first obtain an initial pose model which is then refined iteratively.

6.4 Description of Input Data

Human 3.6M Dataset The dataset we consider for the development of our HPE algorithm is Human 3.6M (Ionescu et al., 2014), which includes about 3.6 million video frames with associated labelled joints and poses of different human subjects performing actions. Relevant for us are the motion capture (MoCap) data (provided as joints rotations and translations) acquired with the Vicon MoCap System; data of 11 subjects performing 15 different actions are available. The 3D joint poses are provided as transformation matrices evaluated with respect to a fixed world origin as described in the next subsection.

Additionally, we consider the corresponding video frames captured from high resolution RGB cameras from 4 different viewpoints. This is done to ensure that we take in consideration a sufficiently varied set of poses captured from different view points. We consider the 4 views of each pose as distinct instances. Furthermore, we are given also the positions of the MoCap skeleton mapped into the image domain. This is used for the 2D joints inference in images, as explained in the following. As in (Zhou et al., 2016), we use 5 subjects (S1, S5, S6, S7, S8) for the training stages, and 2 subjects (S9, S11) for testing. Moreover, we consider only 18 out of the entire set of 32 3D joints by excluding joints corresponding to fingers and toes and by merging together joints corresponding to the same 3D position in order to avoid redundancy in the data. Therefore, for each video frame we have the association among the image, the 3D joint poses, and the 2D joints mapped in the image.

PGA-based Features We now describe the basic principles used for extracting features representing the pose of each group. A MoCap sequence amounts to the poses of a subject at regular time instances. At each time instant the pose of the subject is represented by a given configuration of its joints. In detail, a skeleton $\mathcal{J}$ is specified by 18 joints, where the first one is the index of the root joint. Each joint has a single parent joint, except from the root joint. The configuration of the i-th joint is represented by a homogeneous transformation matrix $T_i \in SE(3)$, a *Lie Group* with identity element defined by the 4×4 identity matrix. By defining a proper metric the Lie Group is a Riemannian manifold, on which we can define (via the exponential mapping) the notion of geodesic between two elements on the manifold (see (Flaherty and do Carmo, 2013; Zefran et al., 1998; Duan et al., 2013)), which is locally

the shortest path that connects two group elements. Henceforth each joint is considered as a rigid body moving in space with respect to some coordinate system. Note that this coordinate system may change according to the MoCap system used for acquiring the data.

We breakdown the skeleton into 11 sub-body groups G_s, with $s = 1, \ldots, 11$. Each group contains M_s joints and is defined as $G_s = \{J_{\psi(1)}, \ldots, J_{\psi(M_s)}\} \subseteq \mathcal{J}$, with $\psi(\cdot)$ providing the relation of the group joint indices with respect to the skeleton indexes. All joints belonging to a group have a parent within the same group, except the root of the group, which is included in at least one other group, whenever it is not the root of the entire skeleton, this proviso is required by the reconstruction of the full-body pose (Algorithm 4.

Table 6.1. Average geodesic distance between the Karcher mean and the rotations of each joints for each group over the whole dataset.

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}	G_{11}
J_1	1.102	1.152	1.152	1.149	1.144	1.143	1.108	1.145	1.106	1.110	1.141
J_2	1.102	1.521	1.521	1.521	1.524	1.518	1.108	1.535	1.106	1.110	1.510
J_3	-	1.520	1.519	1.519	1.540	1.521	-	1.530	-	-	1.519

Breaking down the skeleton into groups is motivated by the idiosyncratic motion of body parts, and to appraise this fact we use the Da Vinci's Vitruvian pose as the reference skeleton configuration, adapting an idea of (Taylor et al., 2012). The Vitruvian pose and the joint groups considered here are shown in Fig. 6.2. Now, given a pose, we find the transformation between the current pose configuration and the Vitruvian pose, for each group G_s, $s = 1, \ldots, 11$. Then, the pose feature set for each group is obtained from the principal direction, computed via *Principal Geodesic Analysis* (Fletcher et al., 2004) from these transformations.

More specifically, for each G_s, $s = 1, \ldots, 11$ the transformation matrices mapping the joints from a current arbitrary pose to the Vitruvian pose are computed, taking into account the dependencies from the parent pose. We compute the Karcher mean (Karcher, 1977) μ of the group transformations, following the algorithm of Afsari (Afsari et al., 2013). In particular, regarding rotation averaging, the center of mass should be within a geodesic distance no larger than $\pi/2$ in order to be unique, and thus well defined (Kendall, 1990; Afsari et al., 2013; Hartley et al., 2013). Table 6.1 shows the average geodesic distance between the intrinsic mean and the rotations of the individual joints for each group over the whole dataset, suggesting that the Karcher mean computation is well defined for this particular choice of groups.

Hence we compute the tangent space of SE(3) at μ and select the principal direction. This direction is the one that best interprets the variability of the motion that the group of joints performs in order to return to the configuration of joints of that sub-body group, in the Vitruvian rest pose. The actual computation of the principal direction in SE(3) is given in (Natola et al.,

2015b), and for the transformation considered here the whole computation is resumed in Algorithm 3.

Algorithm 3: Feature extraction for the pose of a group G_s of joints

Data: The pose of the group G_s given by the corresponding set of homogeneous transformations $\{T_{\psi(1)}, \ldots, T_{\psi(M_s)}\}$; the Vitruvian joints configuration $\{T^V_{\psi(1)}, \ldots, T^V_{\psi(M_s)}\}$.

Result: Feature vector for the pose of the group G_s

1. Move the root of G_s to the root of the corresponding group in Vitruvian pose.
2. Compute the "disparity" between each joint current pose and the Vitruvian pose as $\hat{G}_s = \{\hat{T}_{\psi(1)}, \ldots, \hat{T}_{\psi(M_s)}\}$, taking into account the dependency of each joint pose from its parent pose.
3. Compute the Karcher mean as in (Afsari et al., 2013), extending it to translation.
4. Compute the variance S as in (Fletcher et al., 2004), but using the twist $\mathbf{u}^\vee = (\omega^\top, \mathbf{v}^\top)^\top$, obtained from the Lie algebra of the given transformations, to extend the PGA to SE(3), with ω and $\mathbf{v}$ the instantaneous angular and linear velocities, as in (Natola et al., 2015b).
5. Compute the eigenvector and eigenvalues of S and return the first principal direction in the Lie algebra $se(3)$.
6. Build the feature vector in $\mathbb{R}^7$ using the instantaneous angular and linear velocities from the principal direction, forming a twist, together with the norm of the instantaneous linear velocity (Natola et al., 2015b).

2D joints estimation from Monocular images In both learning and testing stages we extract PHOG visual features for each considered group. For this purpose, given an image sampled from a video of the dataset in Human 3.6M, the first step is the estimation of the 2D joints together with suitable surrounding boxes in the image domain.

In detail, since we have considered the 3D skeleton subdivided into 11 groups we recover 11 boxes (or windows), one for each imaged group. From each of these boxes we extract the most suitable image descriptors for our purpose, that are the Pyramid Histogram of Oriented Gradients (PHOG) (Bosch et al., 2007; Dalal and Triggs, 2005). We have decided to consider a pyramid with levels equal to 0 and 1 and 8 bins spanning an angle of 360 degrees, for each joint in

the group, this choice leads to feature vectors of size $m, m \in \{16, 24, 32\}$.

The estimation of the 2D joints from images is performed using the state-of-the-art approach (Yang and Ramanan, 2013). This approach is particularly suitable for the estimation of the sought-after boxes surrounding joints of human body. We train a model using the algorithm described in (Yang and Ramanan, 2013) using images sampled from the videos in the Human 3.6M dataset. In particular, we used 61750 images for training taken by the 5 different subjects (S1, S5, S6, S7, S8) performing all the actions, provided together with the 2D joints positions. We used 24700 images for testing taken from the remaining subjects (S9, S11) performing the same actions. From the boxes obtained we consider the central points being the 2D joints. Note that we know the ordering of the parts and so of the joints. Fig. 6.4 shows the result of the boxes extraction for two different testing images and the process of PHOG extraction from an image of a group when the PHOG level is set to 0.

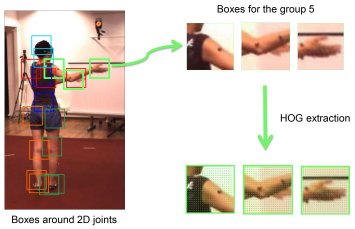


Figure 6.4. **Left**: 2D joints estimation using (Yang and Ramanan, 2013); **Right**: HOG descriptor extraction for a group of joints.

6.5 Features to poses mapping: a hierarchical model

In this section we present the hierarchical model connecting 3D poses and visual features, which make it possible to infer a human pose from the visual features. The hierarchical model takes care of the main aspects of this inference process. First of all it generates a dictionary of poses, for each group. The dictionary collects poses in clusters, where the similarity within a cluster is defined according to the parameters of the underlying distribution. In particular, the dictionary for the poses is a list of indexes specifying for each pose the set of poses sharing the same partition block – or the same parameters. Because a set of similar poses admits several views, the visual features indexed in the same partition generate a mixture of features too. Finally, a principle of compatibility amid clusters of different groups is defined.

In this section we consider $(X_1, X_2, \ldots, X_N)$, $(Y_1, Y_2, \ldots, Y_N)$ sets of real valued random variables; with $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots \mathbf{y}_N)$ their realization. In particular, we consider here a multivariate $\mathbf{X}$, for the principal direction of the poses of a group of joints, such that a random sample of observations $\mathbf{x}_i \in \mathbb{R}^7$. We consider also a multivariate $\mathbf{Y}$ for the PHOG features, with $\mathbf{y}_i \in \mathbb{R}^m$, $m \in \{16, 24, 32\}$. To simplify reading we sometimes

talk about poses, though in fact we consider the twists obtained by the principal direction of the set of rototranslations of the joints of a group, with respect to the same joints in the Vitruvian pose, as explained in Section 6.4.

Given the training sets $D_s^X = \{\mathbf{x}_i | \mathbf{x}_i \in \mathbb{R}^7, i=1,\ldots,N\}$ and $D_s^Y = \{\mathbf{y}_i | \mathbf{y}_i \in \mathbb{R}^m, i=1,\ldots,N, m \in \{16,24,32\}\}$ as the sampled pose and visual features for a group $G_s, s=1,\ldots,11$ (a subset of joints as specified in Fig. 6.2), we want to partition these sets, though neither the partition dimensions nor the specific allocations are known. Hence we resort to the Bayesian nonparametric perspective on mixtures with countably infinite number of components. In this perspective we are given a measurable space $\mathbb{X}$, a discrete measure μ on this space, a collection of continuous observations, latent variables $(\theta_1,\ldots,\theta_K)$ admitting a distribution, with K a random number $\leq N$, and a probability distribution function $F(\cdot|\theta_i)$, parametrized by the random variables θ_i. This setting leads to the popular Dirichlet process mixture model, where $F(\cdot|\theta_i)$ is the kernel of the mixture, here the normal distribution, and $\mu \sim DP(\alpha, \mu_0)$, is the mixing measure, with concentration parameter α and mean $E\{\mu\} = \mu_0$. This is usually expressed in a hierarchical representation as:

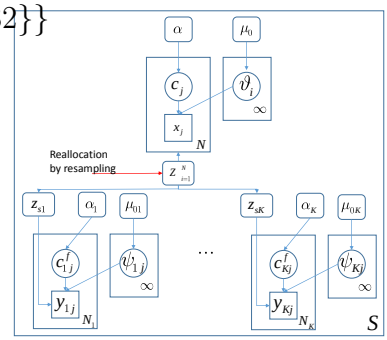


Figure 6.5. Plate representation of $S = 1,\ldots,11$ fold replication of the stacked DPM for pose and visual features. Inner plates are replicated for each DPM.

$$\begin{aligned} &X_i|\theta_i = F(\cdot|\theta_i), && i = 1,\ldots,K \\ &\theta_1,\ldots,\theta_N|\mu \sim_{iid} \mu \text{ and} && \mu \sim DP(\alpha,\mu_0). \end{aligned} \tag{6.1}$$

Then $X \sim \int F(X|\theta)d\mu(\theta)$ is a mixture of distributions with countably infinite number of components (Lo, 1984; Ferguson, 1983). Since the measure μ is discrete, each pair of latent random variables can take the same value with probability $p > 0$. Where the taken value is precisely that of a mixture component. Hence the observations will be allocated by the latent variables to a random number of components.

Different representations have been given of the DPM since (Ferguson, 1973) and several methods have been devised to sample the mixture parameters from the $DP(\alpha,\mu_0)$ (see (Gorür, 2007; Sudderth, 2006)). Recently a number of contributions have explored advanced methods to obtain a parallel implementation (Lovell et al., 2012; Chang and Fisher III, 2013), and to obtain a distribution on the partition of the tangent space to the sphere (Straub et al., 2015), introducing mixture models for data lying on the sphere, and on Riemaniann manifolds (Kim et al., 2015). In this work we did not consider our data as placed on a curved manifold. Despite features data for poses are

obtained from the principal direction on SE(3), each twist extended with the velocity norm, as described in Section 6.4 is independent of the others and forms an exchangeable set. As we do not consider any trajectory between the pose feature vectors we may not consider them on a curved manifold, though we are exploring the interesting modeling that a manifold representation could lead to. Several approaches have also considered different forms of hierarchical and nested NPB models. Though here we could not use the hierarchical model of (Teh et al., 2012), since the pose clusters of the same group, likewise the visual features, do not share any element. Neither could be used across groups, since groups have different ranges of PHOG variates and the number of clusters depends on the number of poses of a specific body part.

Our proposed hierarchical model relies on the hypothesis that for the training datasets there exists an index set $\{Z\}_{i=1}^{N}$, with a bijective mapping h between any two datasets. So, for each PHOG feature vector $\mathbf{y}_i$ there exists a corresponding pose vector $\mathbf{x}_i$ in the training set. This fact does not affects generality nor exchangeability, as we see below, since the index set labels the sampled features not the partitions.

To generate an exchangeable random partition for the mixture of poses, we consider the well known Chinese restaurant process (CRP) (Pitman, 2006). On the other hand, to compute the parameter α we followed the approach of (West, 1992), defining the prior of α as coming from the class of mixtures of gamma distributions, with small initial scale and shape parameters. For inference we resort to Gibbs sampling (Neal, 2000; Jain and Neal, 2004) with conjugate priors.

Given the distribution on the partition induced by the mixture model, a finite set of parameters $\hat{\theta}_1, \ldots, \hat{\theta}_K$ is obtained, together with a cluster indexing $\mathbf{c} = (c_1, \ldots, c_N)$ for each element in the training set. The prediction of a new pose $\mathbf{x}_{N+1}$ is defined by the posterior predictive distribution:

$$p(\mathbf{x}_{N+1}|\mathbf{X}) = \sum_{c_1,\ldots c_{N+1}} \int p(\mathbf{x}_{N+1}|c_{N+1},\theta)p(c_{N+1}|\mathbf{c})p(\mathbf{c},\theta|\mathbf{X})d\theta \tag{6.2}$$

Here:

$$p(\mathbf{c},\theta|\mathbf{X}) = \frac{1}{H}\prod_{k=1}^{K}\mu_0(\theta_k)\prod_{j=1}^{n}F(x_j|\theta_{c_j})P(c_j), \tag{6.3}$$

where H is the marginal likelihood of the mixture of Normals given the computed parameters. And, according to the sampling process induced by the CRP, $p(c_{N+1}|\mathbf{c})$ is:

$$p(c_{N+1} = k|\mathbf{c}) = \begin{cases} \dfrac{n_k}{N-1+\alpha} & k \leq K \\ \dfrac{\alpha}{N-1+\alpha} & \text{otherwise} \end{cases} \tag{6.4}$$

Here n_k is the size of the set of elements in $\mathbf{c}$ having value k. Since poses are continuous and somehow unpredictable, the case that a new pose asks for

the initialization of a new cluster has probability greater than zero. However, once the partition is specified, we make it available to the visual inference, recovering the association between the index set $\{Z\}_{i=1}^{N}$, and each element in each cluster of the dictionary. Because of the label switching problem we prefer to reallocate the indexes $\{Z\}_{i=1}^{N}$ to the clusters. Hence, for each pair $\hat{\theta}_{c_i} = (\eta_{c_i}, \Sigma_{c_i})$ we sample a number of pose vectors $\{\mathbf{u}\}_{|\mathbf{c}_i|}$, proportional to the current ones from $(\eta_{c_i}, \beta D)$, with $\Sigma_{c_i} = UDU^{\top}$, and β a filtering parameter. Given the sampled set we find, in the training set D_s^X, the pose vectors $\mathbf{x}$ which minimize the square error, w.r.t. some specific threshold, i.e. $\{\mathbf{x} \in \mathbf{X} | ||\mathbf{x} - \mathbf{u}||_2 \leq \epsilon, \epsilon > 0\}$. This fact allows, at the same time, to regularize the clusters around their mean, and to reallocate the observations into the clusters together with the observation index set $\{Z\}_{i=1}^{N}$. Therefore according to the model, the induced partition, and the reallocation, given elements $s = \{\mathbf{x}_{s1}, \ldots, \mathbf{x}_{sk}\} | \hat{\theta}_{c_j}$ we have that $h^{-1}(s) = z_{sj}$, a subindex set $z_{sj} \in \{Z\}_{i=1}^{N}$, such that $h(z_{sj}) = \{\mathbf{y}_{s1}, \ldots, \mathbf{y}_{sk}\}$, namely it returns a choice of visual features. The subindex z_{sj} specifies which set of features, having index in $\{Z\}_{i=1}^{N}$ should be allocated to the cluster generated by parameters θ_{c_j}, due to the bijection between the training data. Repeating this for all parameters $\hat{\theta}_{c_j}$, $j = 1, \ldots K$, and for each group, a CRP process is computed for each feature set indexed by z_{sj}. The probability measures generating these new set of DPM, are obviously specific for each PHOG feature set. The structure of the hierarchical model is illustrated in Fig. 6.5. Each feature set indexed by z_{sj} can

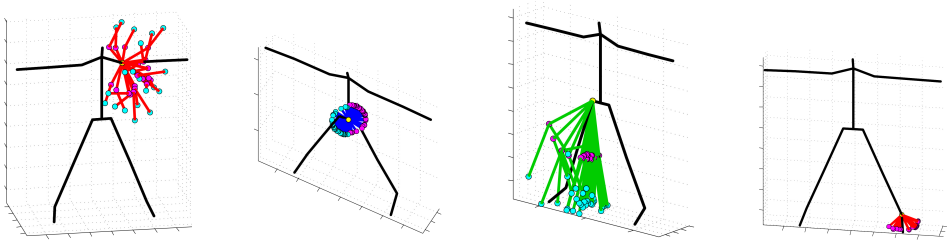


Figure 6.6. Most representative poses of the learned dictionary for the groups *Left Arm, Hips, Right Leg, Left Foot*, with respect to the "Vitruvian pose".

specify different views of the same pose, and possibly under different lighting conditions. Further, we expect that similar poses of different people, yet belong to the same cluster, and the PHOGs might capture this, when represented by a mixture distribution. Thus we induce a new partition exploiting the Gamma additive property. For each cluster of poses, generated by each group, there exists a set of models $\mathbb{M}_s = (\mathcal{M}_{PHOG}^1, \ldots, \mathcal{M}_{PHOG}^{K_s})$, with K varying according to the group $s, s = 1, \ldots, 11$.

Now, given a new observation $\mathbf{y}^{\star}$, this could be either a query or a new measure. Then the posterior predictive of eq. (6.2) should integrate with respect to the parameters of the feature set indexed by z_{sj}, for $j = 1, \ldots, K$ and with respect to each feature set $\mathbf{Y}_{z_{sj}}$, collected in the training. Without

loss of generality we can do this into two steps. In the first step we compute the density, finding the model that best fits $\mathbf{y}^\star$. We can do this because the index set for the visual features is not required for this step:

$$\arg \max_{\mathcal{M}_{PHOG}} p(\mathbf{y}^\star|\xi) = \sum_h \sum_j \pi_{hj} \varphi_h(\mathbf{y}^\star|\xi_{hj}, \mathcal{M}^h_{PHOG}). \tag{6.5}$$

Here the πs are the mixing proportion and $\varphi(\cdot|\xi)$ is the Normal density with parameters ξ for the specific PHOG features set. Once the model is chosen, hence the cluster, the predictive distribution in eq. (6.2), can be applied to the PHOG feature $\mathbf{y}^\star$. Note that if a new component is generated, this now will have its reference pose being the mean of the cluster it is hooked to. Note that if the subindexes of the clusters generated by the visual features $\mathbf{y}$ with subindex z_{sj} are needed, to identify a particular feature and its connection to a particular pose, then a resampling is necessary, as we did with the poses. Otherwise the mean pose can be used. We can see this process as a funnel guiding visual features into the small opening of the pose set, and possibly widening the opening as new observations come in.

Algorithm 4: Consistent pose cluster selection.

Data: Pairwise group compatibility probabilities r_{ij} (eq. 6.6).
Result: Most likely set of consistent pose clusters.

```
Find the most likely pose cluster for the root group (G_8);
Add all the connected groups of G_8 (denoted children(G_8)) in the set G_open;
while G_open is not empty do
    for Each group G_s ∈ G_open do
        Find its most likely pose cluster taking into account the
        compatibilities {r_ij}_{i∈{1,M_s}} with respect to the selected cluster
        j of its parent group parent(G_s)
    Remove (G_s) from G_open;
    Add children(G_s) in G_open
```

The final inference step requires a principle of compatibility amid groups from which derive the consistent pose selection summarized in Algorithm 4. We define the intergroup clusters compatibility as follows. Let i, j, be two clusters from groups q and s. Let $W_{ij} = |z_{qj} \cap z_{si}|$ with $|\cdot|$ the cardinality and let $D_{ij} = z_{qj} \cup z_{si}$ and $p(m_{ij} = 1) = W_{ij}/|D_{ij}|$.

The probability that the two intergroup clusters are compatible is given as:

$$r_{ij} = \frac{p(D_{ij}|m_{ij} = 1)p(m_{ij} = 1)}{p(D_{ij}|m_{ij} = 1)p(m_{ij} = 1) + p(D_{ij}|m_{ij} = 0)(1 - p(m_{ij} = 1))} \tag{6.6}$$

With

$$p(D_{ij}|m_{ij}=1) = \gamma \sum_{D_{ij}} \pi_i \delta_{D_{ij}}(\mathbf{x}) + (1-\gamma) \sum_{D_{ij}} \pi_j \delta_{D_{ij}}(\mathbf{x}) \tag{6.7}$$

Where $\delta_{D_{ij}}(\mathbf{x}) = 1$ if $\mathbf{x} \in D_{ij}$ and zero otherwise, π_i and π_j are the mixing proportions of the DPM of the two clusters, and $0 \leq \gamma \leq 1$ balances the contribution from the two clusters. While, where the two clusters are completely uncorrelated:

$$p(D_{ij}|m_{ij}=0) = \prod_{D_{ij}} \pi_i \pi_j \tag{6.8}$$

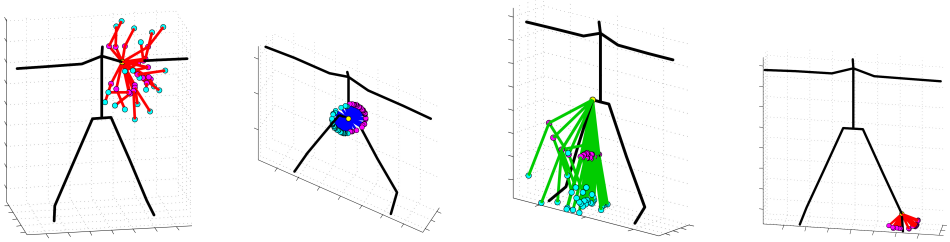


Figure 6.7. Most representative poses of the learned dictionary for the groups *Left Arm, Hips, Right Leg, Left Foot*, with respect to the "Vitruvian pose".

Full-pose consistency In order to reassemble the full-body pose of the subject we consider the structure of the skeleton and the proposed decomposition into groups. Starting from the group of the hips, which contains the root joint, we seek the most likely entries of the pose dictionary for the connected groups (groups which share at least one joint). Let us indicate with η_{G_s} the pose (cluster) that best fits the image query specified by $(\mathbf{y}^{\star}_{G_s})$, for each group G_s. We seek to maximize $p((\eta_{G_1}, \ldots, \eta_{G_{11}})|(\mathbf{y}^{\star}_{G_1}, \ldots, \mathbf{y}^{\star}_{G_{11}}))$. The tree-like structure of the human skeleton, allows for a formulation of the problem using graphical models and the optimal solution is easily obtained using the max product algorithm. Once the most likely consistent set of pose clusters has been selected, the full-body pose is reconstructed. In particular, starting from the root and going toward the extremities, we obtain the most representative pose of the selected cluster for each group and we make the reference frames of the shared joints between the clusters coincide.

6.6 Results

Dictionary learning As described in Section 6.4, we consider the dataset Human 3.6M (Ionescu et al., 2014) to evaluate our 3D pose estimation algorithm. In order to obtain the dictionaries of the 3D poses we first apply the decomposition of the joints in groups according to Fig. 6.2 and then compute PGA-based features for each group joints, as described in Section 6.4. As

the dataset contains 3D poses synchronized with video frames at a high rate (50 Hz), we subsample with a factor of 5 in order to remove redundant data. Further we compute the PHOG features as described in Section 6.4. The number of clusters generated for each group by the DPM models are reported in Table 6.2.

Table 6.2. Number of clusters generated by the DPM models for the PHOG and the PGA-based features for each group of joints.

Groups	1	2	3	4	5	6	7	8	9	10	11	12
Nr. of pose clusters	56	155	38	85	20	49	90	88	58	49	52	16
Avg. nr. of visual components	18	31	31	25	22	22	4	22	22	11	18	13

The significance of pose clusters is shown in Fig. 6.7, where the mean poses are visualized for the groups *Left Arm, Hips, Right Leg, Left Foot.*

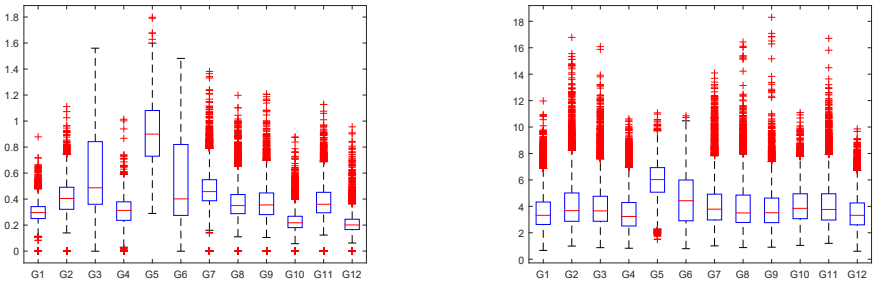


Figure 6.8. Error distribution for the PHOG (left) and the PGA (right) features.

3D pose estimation Using the learned dictionary of poses and visual features we perform 3D pose estimation for the testing part of the dataset, namely for the actions performed by subjects S9 and S11. For each query image, the 2D joint positions in the image are estimated by using (Yang and Ramanan, 2013), and they are grouped together forming the groups of Fig. 6.2. For each group, the PHOG features are then extracted, as described in Section 6.4, and the corresponding cluster of the visual dictionary is selected as the most likely one according to the learned hierarchical model. We calculate the error of the visual features as the euclidean distance of the extracted features with respect to the most representative visual features of the selected cluster. The mean of this error together with the 25th and 75th percentiles for each group, are shown in the left box-plot of Fig. 6.8. Note that as the errors refer to distances, we expect that they follow a χ^2 distribution instead of a normal one. We observe that the errors of the PHOG features are low in average for most of the groups. The groups corresponding to the hands and the arms (G_3, G_4, G_5, G_6) show higher errors, mainly because of the high variability of their appearance.

The 3D pose of the whole body is obtained according to Algorithm 4.

[1]More results are reported in the supplementary material

Table 6.3. Average per joint error between the estimated 3D pose and the ground truth in mm. Best values in bold.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
LinKDE(Ionescu et al., 2014)	132.71	183.55	132.37	164.39	162.12	205.94	150.61	171.31
Li et. al(Li and Chan, 2014)	-	136.88	96.94	124.74	-	168.68	-	-
Tekin et al.(Tekin et al., 2015)	102.39	158.52	87.95	126.83	118.37	185.02	114.69	107.61
Zhou et al.(Zhou et al., 2016)	87.36	109.31	**87.05**	103.16	116.18	143.32	106.88	**99.78**
Ours	**48.82**	**56.31**	95.98	**84.78**	**96.47**	**105.58**	**66.30**	107.41
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	**Average**
LinKDE(Ionescu et al., 2014)	151.57	243.03	162.14	170.69	177.13	96.60	127.88	162.14
Li et. al(Li and Chan, 2014)	-	-	-	-	132.17	69.97	-	-
Tekin et al.(Tekin et al., 2015)	136.15	205.65	118.21	146.66	128.11	**65.86**	**77.21**	125.28
Zhou et al.(Zhou et al., 2016)	124.52	199.23	107.42	118.09	**114.23**	79.39	97.70	113.01
Ours	**116.89**	**129.63**	**97.84**	**65.94**	130.46	92.58	102.21	**93.15**

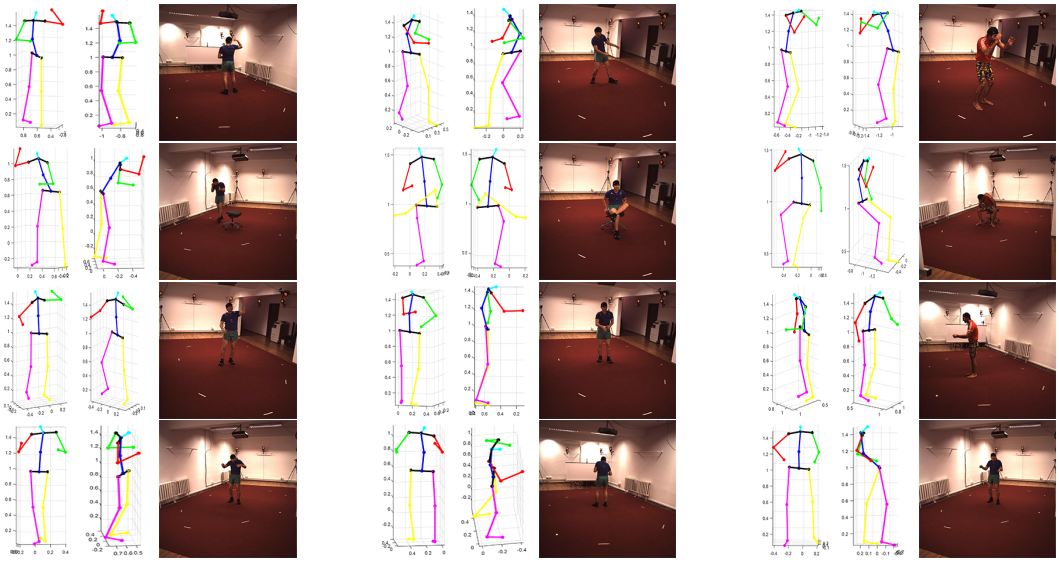


Figure 6.9. Examples of query images and the recovered 3D pose [1].

Examples of the recovered poses for query images of the subjects S9 and S11 are shown in Fig. 6.9. We calculate the euclidean distance of the PGA-based features of the true 3D pose of the subject, with respect to the most representative PGA-based features of the selected cluster for each group. The mean distance for each group together with the 25th and 75th percentiles, are shown in the right box-plot of Fig. 6.8. We note that the average errors of the PGA-based features are small for all groups, apart from G_5 and G_6 which correspond to the right arm and the right hand. The fact that the PGA features reside in a deeper level of the hierarchical model affects the presence of an increased number of errors above the 95th percentile.

We also compute the mean error of the joint positions of the recovered 3D pose with respect to the ground truth 3D pose of the subject. This error, compared to the error of other state of the art approaches is reported in Table 6.3. The results show that our method gives slightly worse results only with respect to (Zhou et al., 2016) for the 'Eating' and 'Purchases' actions, and for the walking actions with respect to (Tekin et al., 2015) and (Zhou

et al., 2016). In summary, the proposed method outperforms other recently proposed state of the art 3D pose estimation methods both in average and also for the vast majority of actions considered in the Human 3.6M dataset.

Efficiency of the method For the 2D joints estimation training uses 61750 frames of the Human 3.6M dataset taking about 10^4 sec., (Yang and Ramanan, 2013) does not report efficiency. For the hierarchical DPM we consider a training set of 130272 frames, asking for $\sim 8.5 \times 10^5$ seconds for the poses partitioning and $\sim 7 \times 10^4$ seconds for the visual features partitioning. This considering main Gibbs cycles of 1800 iterations. Full-pose consistency takes around 0.05 seconds for a single query, and the total percentage of queries not satisfying it are around 23%. Once parameters are learned pose computation takes around 0.96 seconds, with PGA and group computation taking around 0.07 seconds. These results are obtained with a computer equipped with four Xeon E5-2643, 3.70GHz CPUs and 64GB RAM.

6.7 Conclusions

We present a novel method for 3D human pose estimation from a single image based on a hierarchical Bayesian non-parametric model. The proposed model captures idiosyncratic variations of the motion and the appearance of different body parts, identified by groups of joints. The decomposition in groups avoids redundant configurations, obtaining a more concise dictionary of poses and visual appearances. Given the learned model a 3D pose query can be resolved in real-time. The results show that the proposed model is able to generalize and accurately reconstruct the 3D pose of previously unseen subjects. Our results improve the current state of the art though we aim to further ameliorate them, by considering additional constraints of the pose structure. We shall also consider to move the NBP on the Riemann manifold of the features.

Chapter 7

Discovery and recognition of motion primitives in human activities

Marta Sanzari[1], Valsamis Ntouskos[1], & Fiora Pirri[1]

[1] ALCOR LAB, Dipertimento di Ingegneria Informatica Automatica e Gestionale, Sapienza University of Rome

Published: In PloS one, 14.4: e0214499, 2019.

Statement of Contributions of Joint Authorship

Marta Sanzari (Candidate):
Writing and compilation of manuscript, established methodology, data analysis, preparation of tables and figures.
Valsamis Ntouskos (Research Colleague):
Writing and compilation of manuscript, established methodology, data analysis, preparation of tables and figures, co-author of manuscript.

Fiora Pirri (Principal Supervisor):
Supervised and assisted with manuscript compilation, editing and co-author of manuscript.

Here we present the work done for the discovery and recognition of human motion primitives, published at the journal PloS one in 2019. This work introduces a framework for automatically discovering and recognizing human motion primitives from video sequences, introducing the motion flux method. A hierarchical model for the classification and recognition of the unlabeled discovered primitives, for each skeleton group, was built. It was shown that

each primitive category naturally corresponds to movements described using biomechanical terms. Motion primitives categories were proven to be discriminative for characterizing the activity performed by the human subjects in videos, describing an application to abnormal behaviors detection. Finally, a dataset of motion primitives was made publicly available to further encourage result reproducibility and benchmarking of methods dealing with the discovery and recognition of human motion primitives. The dataset can be found at `https://github.com/alcor-lab/MotionPrimitives`.

This Chapter is an exact copy of the journal paper referred to above.

7.1 Abstract

We present a novel framework for the automatic discovery and recognition of motion primitives in videos of human activities. Given the 3D pose of a human in a video, human motion primitives are discovered by optimizing the 'motion flux', a quantity which captures the motion variation of a group of skeletal joints. A normalization of the primitives is proposed in order to make them invariant with respect to a subject anatomical variations and data sampling rate. The discovered primitives are unknown and unlabeled and are unsupervisedly collected into classes via a hierarchical non-parametric Bayes mixture model. Once classes are determined and labeled they are further analyzed for establishing models for recognizing discovered primitives. Each primitive model is defined by a set of learned parameters. Given new video data and given the estimated pose of the subject appearing on the video, the motion is segmented into primitives, which are recognized with a probability given according to the parameters of the learned models. Using our framework we build a publicly available dataset of human motion primitives, using sequences taken from well-known motion capture datasets. We expect that our framework, by providing an objective way for discovering and categorizing human motion, will be a useful tool in numerous research fields including video analysis, human inspired motion generation, learning by demonstration, intuitive human-robot interaction, and human behavior analysis.

7.2 Introduction

Activity recognition is widely acknowledged as a core topic in computer vision, witness the huge amount of research done in recent years spanning a wide

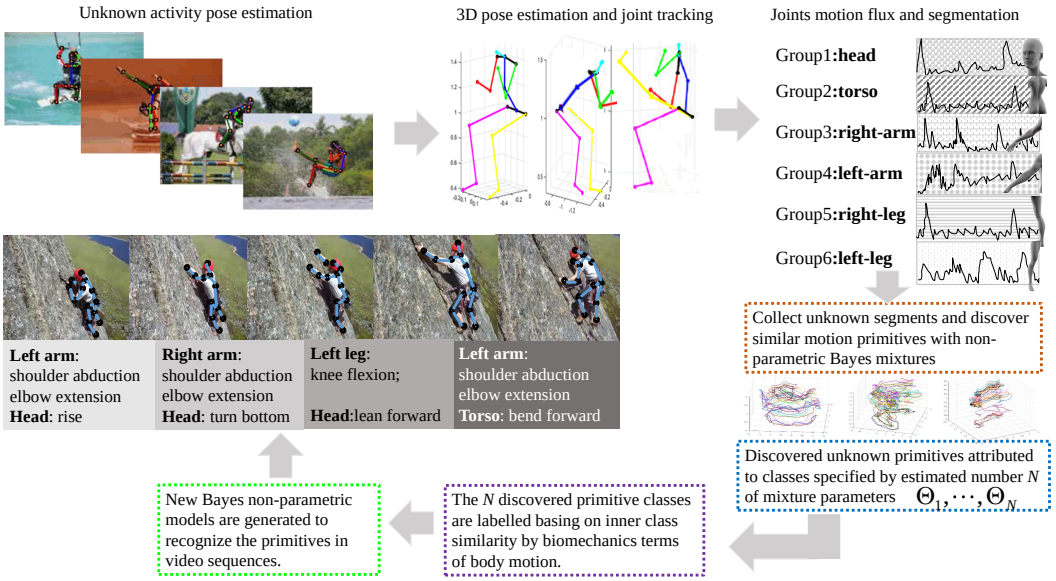


Figure 7.1. The above schema presents the proposed framework and the process to obtain from video sequences the discovered motion primitives.

number of applications from sport to cinema, from human robot interaction to security and rehabilitation.

Activity recognition has evolved from earlier focus on action recognition and gesture recognition. The main difference being that activity recognition is completely general as it concerns any kind of human activity, which can last few seconds or minutes or hours, from daily activities such as cooking, self-care, talking at the phone, cleaning a room, up to sports or recreation such as playing basketball or fishing. Nowadays there are a number of publicly available datasets dedicated to the collection of any kind of human activity, likewise a number of challenges (see for example the ActivityNet challenge (Ghanem et al., 2017)).

On the other hand, the interest in motion primitives is due to the fact that they are essential for deploying an activity. Think about sport activities, or cooking, or performing arts, which require to purposefully select a specific sequences of movements. Likewise daily activities such as cleaning, or cooking, or washing the dishes or preparing the table require precise motion sequences to accomplish the task. Indeed, the compositional nature of human activities, under body and kinematics constraints, has attracted the interest of many research areas such as in computer vision (Yang et al., 2013b; Holte et al., 2010), in neurophysiology (Flash and Hochner, 2005; Polyakov, 2017), in sports and rehabilitation (Ting et al., 2015), and in biomechanics (Hogan and Sternad, 2012) and in robotics (Amor et al., 2014; Moro et al., 2012; Azad et al., 2007).

The goal of this work is to automatically discover the start and end points where primitives of 6 identified body parts occur throughout the course of an activity, and recognize each of the occurred primitives. The idea is that

these primitives sort out a non-complete set of human movements, which combined together can form a wide range of human activities, in so providing a compositional approach to the analysis of human activities.

The steps of the proposed method are as follows. Given a video of a human activity both the 2D pose and 3D pose of the human are estimated (see (Sanzari et al., 2016), and also (Tome et al., 2017)). Once the 3D poses of the joints of interest are determined, we compute the *motion flux*. The motion flux method provides a model from first principles for human motion primitives, and it effectively discovers where primitives begin and end on human activity motion trajectories.

Motion primitives discovered by the motion flux are unknown: they are segments of motion about which only the involved specific body part is known. These primitives are collected into classes by a non-parametric Bayes model, namely the Dirichlet process mixture model (DPM), which gives the freedom to not choose the number of mixture components. By suitably eliminating very small clusters it turns out that discovered primitives can be collected into 69 classes (see Fig. 7.12). For each of them the mixture model returns a parameter set identifying the precise primitive class. We label the computed parameters with terms taken from the biomechanics of human motion, by inspecting only a representative primitive for each discovered class. Out of these generated classes we form a new layer of the hierarchical model, to generate the parameters for each class, further used for primitives recognition. Under this last models each primitive category is approximated by a DPM with a number of components mirroring the inner idiosyncratic behavior of each primitive class.

Motion primitives classification is finalized by providing a label for each primitive. Namely, given an activity (possibly unknown) and an unknown primitive discovered by motion flux, we find the model the primitive belongs to, hence the primitive is labeled by that model.

Experiments show that the motion flux is a good model for segmenting the motion of body parts. Likewise, the unsupervised non-parametric model provides both a good classification of similar motion primitives and a good estimation of primitive labels, as shown in the results (see Section 7.7). The approach therefore is quite general and it turns out to be very useful to any researcher who would like to explore the compositional nature of any activity, using both the proposed method and the motion primitives dataset provided.

To the best of our knowledge just few works, among which we recall (Yang et al., 2013b; Holte et al., 2010), have faced the problem of discovering motion primitives in video activities or motion capture (MoCap) sequences, quantitatively evaluating the ability to recognize them.

Despite the lack of works on motion primitives we show that they are quite an expressive *language* for ascertaining specific human behaviors. To prove that, in a final application for video surveillance, described in Section 7.8, we show that motion primitives can play a compelling role in detecting distinct

classes of dangerous activities. In particular, we show that dangerous activities can be detected with off-the- shelf classifiers, once motion primitives have been extracted in the videos. Comparisons with state of the art results prove the relevance of motion primitives in discovering specific behaviors, since motion primitives embed significant time-space features easily usable for classification.

The contributions of the work, schematically shown in Fig. 7.1 are the followings:
1. We introduce the motion flux method to discover motion primitives, relying on the variation of the velocity of a group of joints.
2. We introduce a hierarchical model for the classification and recognition of the unlabeled primitives, discovered by the motion flux.
3. We show a relevant application of human motion primitives for video surveillance.
4. We created a new dataset of human motion primitives from three public MoCap datasets ((Ionescu et al., 2014), (CMU,), (Mandery et al., 2015)).

7.3 Related work

Human motion primitives are investigated in several research areas, from neurophysiology to vision to robotics and biomechanics. Clearly, any methodology has to deal with the vision process, and many of the earliest more relevant approaches to human motion highlighted that understanding human motion requires view independent representations (Weinland et al., 2006; Li et al., 2010) and that a fine grained analysis of the motion field is paramount to identify primitives of motion. In early days this required a massive effort in visual analysis (Turaga et al., 2008) to obtain the poses, the low level features, and segmentation. Nowadays, scientific and technological advances have made it possible to exploit several methods to measure human motion, such as the availability of a number of MoCap databases (Ionescu et al., 2014; Mandery et al., 2015; Sigal et al., 2009), see for a review (Moeslund et al., 2006). Furthermore recent findings result in methods that can deliver 3D human poses from videos if not even from single frames (Akhter and Black, 2015b; Sanzari et al., 2016; Zhou et al., 2016; Tome et al., 2017). Since then 3D MoCap data have been widely used to study and understand human motion, see for example (Ntouskos et al., 2013; Ntouskos et al., 2015a; Pirri and Pizzoli, 2011) in which Gaussian Process Latent Variable Models or Dirichlet processes are used to classify actions, or (Natola et al., 2015b) in which a non-parametric Bayesian approach is used to generate behaviors for body parts and classify actions based on these behaviors. In (Natola et al., 2015a) temporal segmentation of collaborative activities is examined, or in (Fanello et al., 2010) different descriptors are exploited to achieve arm-hand action recognition.

Neurophysiology Neurophysiology studies on motion primitives (Bizzi and Mussa-Ivaldi, 1995; Flash and Hochner, 2005; Flash et al., 2013; Viviani and Flash, 1995; Flash and Handzel, 2007; Biess et al., 2007) are based on the idea that kinetic energy and muscular activity are optimized in order to conserve energy. In these works it has been observed that curvature and velocity of joint motion are related. Earliest works such as Lacquaniti et al. (Lacquaniti et al., 1983) proposed a relation between curvature and angular velocity. In particular, using their notation, letting C be the curvature and A the angular velocity, they called the equation $A = KC^{\frac{2}{3}}$ the Two-Thirds Power law, valid for certain class of two-dimensional movements. Viviani and Schneider (Viviani and Schneider, 1991) formulated an extension of this law, relating the radius of curvature R at any point s along the trajectory with the corresponding tangential velocity V, in their notation:

$$V(s) = K(s) \left(\frac{R(s)}{1 + \alpha R(s)} \right)^{\beta} \tag{7.1}$$

where the constants $\alpha \geq 0$, $K(s) \geq 0$ and β has a value close to $= \frac{1}{3}$. An equivalent Power law for trajectories in 3D space is introduced by (Maoz and Flash, 2014) and it is called the curvature-torsion power law and is defined as $\nu = \alpha \kappa^{\beta} |\tau|^{\gamma}$, where κ is the curvature of the trajectory, τ the torsion, ν the spatial movement speed, β and γ are constants.

Computer Vision The interpretation of motion primitives as simple individual actions or gestures is often purported, in any case they are related to segmentation of videos and 3D motion capture data. Many approaches explore video sequences segmentation to align similar action behaviors (Gong et al., 2014) or for spatio-temporal annotation as in (Lillo et al., 2016). Lu et al. (Lu et al., 2015) propose to use a hierarchical Markov Random Field model to automatically segment human action boundaries in videos. Similarly, (Bouchard and Badler, 2007) develop a motion capture segmentation method. n-grams have been used to achieve action recognition based on action primitives.

Besides these works, only (Vecchio et al., 2003; Yang et al., 2013b; Holte et al., 2010; Endres et al., 2013) have targeted motion primitives, to the best of our knowledge. (Vecchio et al., 2003) focuses on 2D primitives for drawing, on the other hand (Yang et al., 2013b) does not consider 3D data and generate the motion field considering Lukas-Kanade optical flow for which Gaussian mixture models are learned. None of these approaches provide quantitative results for motion primitives, but only for action primitives, which makes their method not directly comparable with ours. (Holte et al., 2010; Endres et al., 2013) use 3D data and explicitly mention motion primitives, providing quantitative results. The authors account for the velocity field via optical flow basing the recognition of motion primitives on harmonic motion context descriptors. Since (Holte et al., 2010) deal only with upper torso gestures we compare with

them only the primitives they mention. In (Endres et al., 2013) the authors achieve motion primitives segmentation from wrist trajectories of sign language gestures, obtaining unsupervised segmentation with Bayesian Binning. Again here no comparison for motion primitives discovery or recognition is possible as original data are not available.

Robotics In robotics the paradigm of transferring human motion primitives to robot movements is paramount for imitation learning and, more recently to implement human-robot collaboration (Ijspeert et al., 2013). A good amount of research in robotics has approached primitives in terms of Dynamic Movement Primitives (DMP) (Ijspeert et al., 2013) to model elementary motor behaviors as attractor systems, representing them with differential equations. Typical applications are learning by imitation or learning from demonstration (Gams et al., 2016; Pastor et al., 2009; Kober and Peters, 2009; Park et al., 2008), learning task specifications (Ureche et al., 2015), modeling interaction primitives (Amor et al., 2014). Motion primitives are represented either via Hidden Markov models or Gaussian Mixture Models (GMM). (Asfour et al., 2006) present an approach based on HMM for imitation learning of arm movements, and (Luo and Berenson, 2015) model arm motion primitives via GMM.

It is apparent that in most of the approaches motion primitives are only observed and modeled, instead we are able to learn and model them using respectively the *motion flux* quantity and a hierarchical model. The main contribution of our work is indeed the introduction of a new ability for a robot to automatically discover motion primitives observing 3D joints raw pose data. The outcome of our approach is also a motion primitives dataset not requiring human manual operation.

Our view of motion primitive shares the hypothesis of energy minimality during motion, fostered by neurophysiology, likewise the idea to characterize movements using the proper geometric properties of the skeleton joints space motion. However, for primitive discovery, we go beyond these approaches capturing the variation of the velocity of a group of joints using this as the baseline for computing the change in motion by maximizing the motion flux.

7.4 Preliminaries

The 3D pose of a subject, as she appears in each frame of a video presenting a human activity, is inferred according to the method introduced in (Sanzari et al., 2016). Other methods for inferring the 3D pose of a subject are available, we refer in particular, to the method introduced by (Tome et al., 2017), which improves (Sanzari et al., 2016) in accuracy.

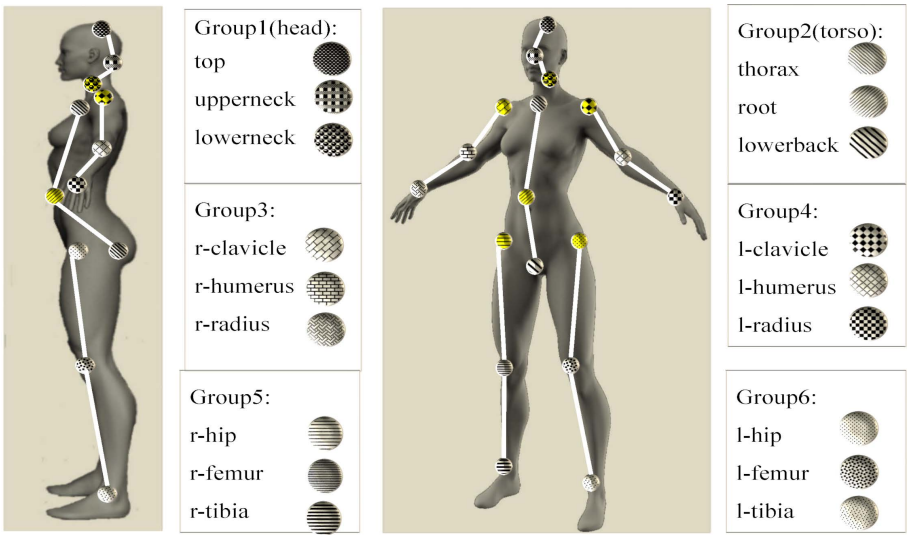


Figure 7.2. The six groups partitioning the human body with respect to motion primitives are shown, together with the joints specifying each group and the skeleton hierarchy inside each group: joints in yellow are the *parent joints* in the skeleton hierarchy.

3D pose data for a single subject are given by the joints configuration. Joints are associated with the subject skeleton as shown in Fig. 7.2 and are expressed via transformation matrices $\mathcal{T}$ in $SE(3)$:

$$\mathcal{T} = \begin{bmatrix} R & \mathbf{d} \\ \mathbf{0}^{1\times 3} & 1 \end{bmatrix} \tag{7.2}$$

Here $R \in SO(3)$ is the rotation matrix, and $\mathbf{d} \in \mathbb{R}^3$ is the translation vector. $\mathcal{T} \in SE(3)$ has 6 DOF and it is used to describe the pose of the moving body with respect to the world inertial frame. $SO(3)$ and $SE(3)$ are Lie groups and their identity elements are the 3×3 and 4×4 identity matrices, respectively. We consider an ordered list $\mathcal{J} = \{j_1^1, j_2^1, \ldots, j_{K-1}^m, j_K^m\}$ of $K = 18$ joints forming the skeleton hierarchy, as shown in Fig. 7.2, with $m = 1, \ldots, 6$ being the groups each joint belongs to. The 6 groups $G_1, \ldots, G_6$ we consider here correspond to head, torso, right and left arm, right and left leg.

Each joint j_i^m, $i = 1, \ldots, 18$, belonging to a group G_m, $m = 1, \ldots, 6$, has one parent joint $j_i^{m,\star}$, which is the joint of the group closest to the root joint $root = j_4^2 \in \mathcal{J}$, according to the skeleton hierarchy, namely it is the fourth joint in the ordered list $\mathcal{J}$ and it belongs to the group G_2, the torso. Parent joints for each group are illustrated in yellow on the woman body in the left of Fig. 7.2, they are in the order $(j_3^1, j_4^2, j_7^3, j_{10}^4, j_{13}^5, j_{16}^6)$.

A MoCap sequence of length N is formed by a sequence of frames of poses. Each frame of poses is defined by a set of transformations $\{\mathcal{T}_{i,m}^k \in SE(3) : k = 1, \ldots, N, m = 1, \ldots, 6\}$ involving all joints $j_i^m \in \mathcal{J}$, $i = 1, \ldots, 18$, according to the skeleton hierarchy. Given a MoCap sequence of length N, for each frame k

the pose of each joint is *root-sequence* normalized, to ensure pose invariance with respect to a common reference system of the whole skeleton. Let $\mathcal{T}^k_{i,m}$ be the pose of the joint j^m_i, according to the skeleton hierarchy, at frame k in the sequence, and let $j^{m,\star}_i$ be the parent node of j^m_i, then the *root-sequence* normalization is defined as follows:

$$\hat{\mathcal{T}}^k_{i,m} = \left((\mathcal{T}^1_{root,2})^{-1}\mathcal{T}^1_{j^{m,\star}_i,m}\right)\left((\mathcal{T}^k_{j^{m,\star}_i,m})^{-1}\mathcal{T}^k_{i,m}\right). \tag{7.3}$$

Here $(\mathcal{T}_{root,2})$ is the transformation of the root node, which is the joint j^2_4 belonging to the group G_2, the torso. Equation (7.3) says that the pose $\mathcal{T}^k_{i,m}$ of joint $j^m_i \in G_m$ at frame k is *root-sequence* normalized if obtained by a sequence of transformations seeing first a transformation with respect to its parent node $(\mathcal{T}^k_{j^{m,\star}_i,m})^{-1}$, at frame k, and then with respect to the transformation of the parent node with respect to the root node, taken at the initial frame of the sequence. In Fig. 7.3 are shown joints position data for each skeleton group after *sequence-root* normalization for all sequences in the dataset. More details on the skeleton structure and its transformations can be found in (Natola et al., 2015b; Sanzari et al., 2016).

7.5 Motion Primitive Discovery

We are considering now the problem of discovering motion primitives within a motion sequence displaying an activity in a video. We begin by providing the definition of a joint trajectory on which the temporal analysis is performed.

Definition 7.5.1 (Joint Trajectory)**.** The trajectory of a joint j is given by the path followed by the skeletal joint j in a given interval of time $I = [t_1, t_2]$. Formally:

$$\xi_j : I \subset \mathbb{R} \mapsto \mathbb{R}^3, \tag{7.4}$$

Based on the definition above, motion primitives correspond to segments of the joint trajectories of a group G. We identify motion primitives as trajectory segments where the variation of the velocity of the joints is maximal and where the endpoints of the segment correspond to stationary poses of the subject (Marr and Vaina, 1982).

Preprocessing To overcome problems related to the finite sampling frequency of the poses in the data, we compute smooth versions of the joint trajectories by cubic spline interpolation. This interpolation provides a continuous-time trajectory for all the joints of the group with smooth velocity and continuous acceleration, satisfying natural constraints of human motion.

Motion Flux The motion flux captures the variation of the velocity of a group with respect to its rest pose. The total variation of the joint group velocity is evaluated along a direction $\mathbf{g}$ that corresponds to stationary poses of the group. For groups 1, 3 and 4 this direction is defined by the segment connecting the

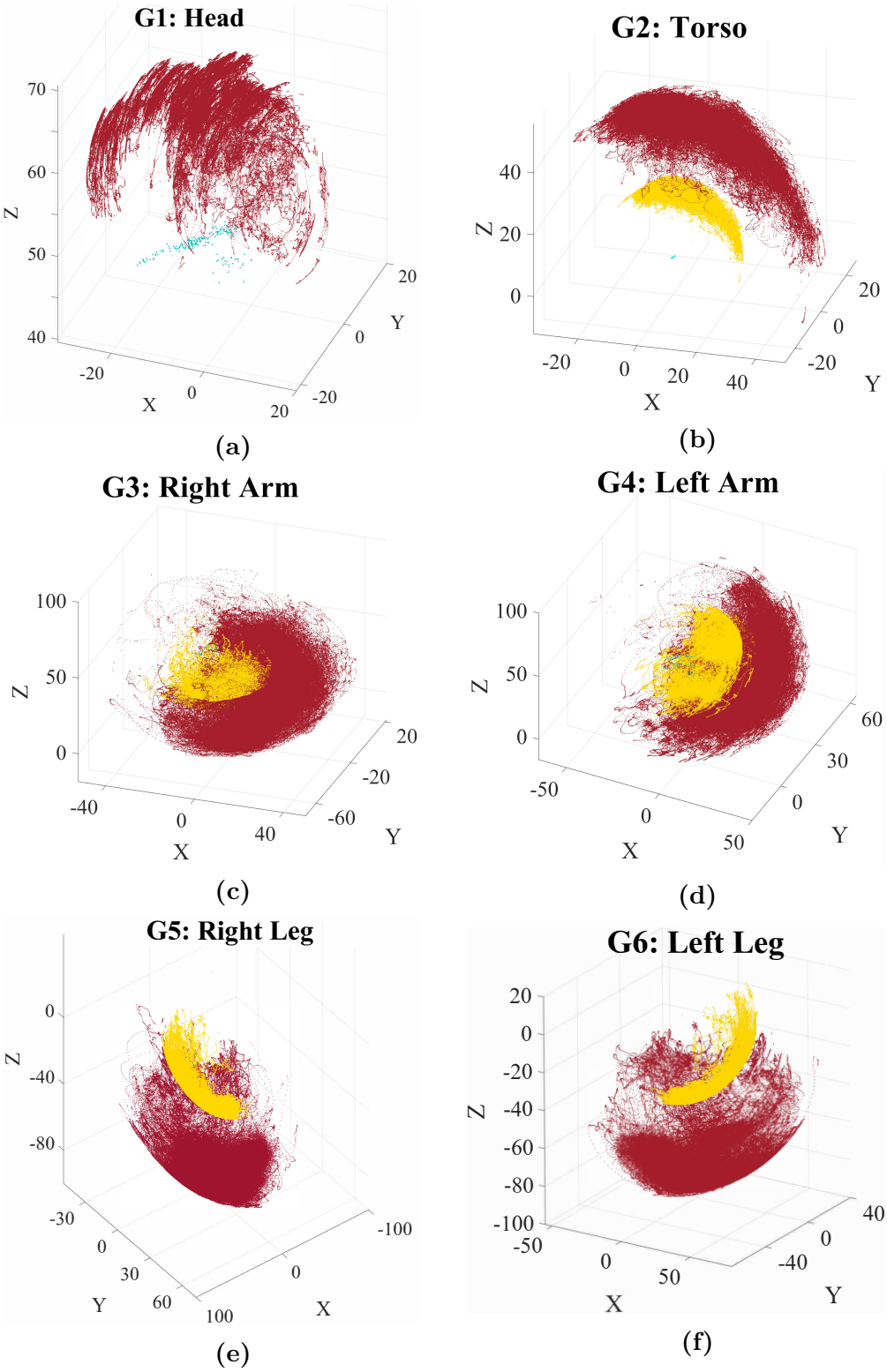


Figure 7.3. Sequences of joint positions, for each skeleton group, after the *root-sequence* normalization described in Section 7.4. Position data are in cm. The green points show the most internal group joint data (e.g. the hip for the leg); the yellow points show the intermediate group joint data (e.g. the knee for the leg); the red points show the most external group joint data (e.g. the ankle for the leg). The joints data are collected from the datasets described in Section 7.7.

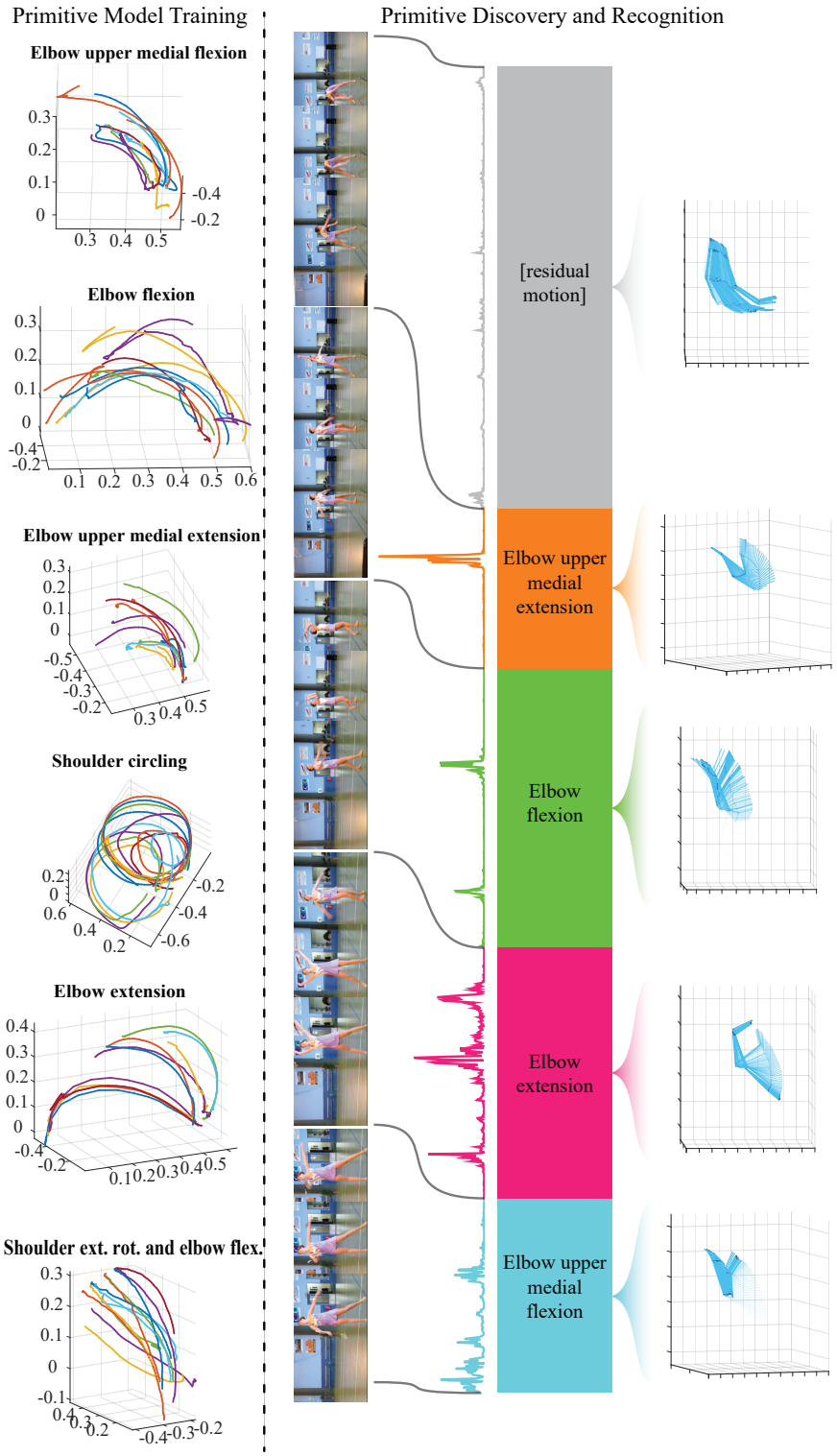


Figure 7.4. Overview of motion primitive discovery and recognition framework. The top section shows primitives of the group 'Arm' from six different categories. Primitives are discovered by maximizing the *motion flux* energy function, presented here above the colored bar, though deprived of velocity and length components. These sets of primitives are used to train the hierarchical models for each category. Primitives are then recognized according to the learned models. The recognized motion primitive categories are depicted with different colors. At the bottom, the group motion in the corresponding interval is shown.

'lowerneck' and 'upperneck' joints while for groups 2, 5 and 6 by the segment connecting the 'root' with the 'lowerback' joints.

Definition 7.5.2 (Motion Flux)**.** Let $G = \{j_1, \ldots, j_K\}$ be a group consisting of K joints and $\mathbf{v}_j$ the velocity of joint $j \in G$. The *motion flux* with respect to the time interval $I = [t_1, t_2]$ is defined as

$$\Phi(t_2, t_1) \doteq \sum_{j \in G} \int_{t_1}^{t_2} |\dot{\mathbf{v}}_j(t) \cdot \mathbf{g}| \ dt. \tag{7.5}$$

Discovery In order to discover a motion primitive, we identify a time interval between two time instances (endpoints) where the group velocity is minimal while the motion flux within the interval is maximal. This is done by performing an optimization based on the motion flux of a group G, as defined in eq. (7.5). More specifically, the time interval of a motion primitive is identified by maximizing the following energy-like function:

$$P(\rho; t_0) \doteq \ \Phi(\rho, t_0) - \frac{\beta_v}{2} \sum_{j \in G} \left(\mathbf{n}\mathbf{v}_j(\rho)^2 + \mathbf{n}\mathbf{v}_j(t_0)^2 \right) + \beta_s \sum_{j \in G} (s_j(\rho) - s_j(t_0)), \tag{7.6}$$

where $s_j(t) = \int_0^t \|\dot{\xi}_j(\tau)\| d\tau$ is the arc length function of ξ_j. The last term of eq. (7.6) is a regularizer based on the length of the trajectory segment, introduced in order to avoid excessively long primitives. The hyper-parameter β_v acts as penalizer associated to the soft-constraint on the stationarity of the poses at the start and end of the primitive, while β_s controls the strength of the regularization on the primitive length. Both β_v and β_s depend on the scaling of the data and the sampling rate of the joint trajectories.

Given a starting time instant t_0, a motion primitive is extracted by identifying the time instant ρ, which corresponds to a local maximum of (7.6). The optimality condition of (7.6) gives:

$$\sum_{j \in G} \left(|\dot{\mathbf{v}}_j(\rho) \cdot \mathbf{g}| - \beta_v \frac{\dot{\mathbf{v}}_j(\rho)}{\mathbf{n}\mathbf{v}_j(\rho)} - \beta_s \|\dot{\xi}_j(\rho)\| \right) = 0. \tag{7.7}$$

Given the one-dimensional nature of the problem, finding the zeros of (7.7) and verifying whether they correspond to local maxima of (7.6) is trivial.

Based on the previous we provide a formal definition of a motion primitive.

Definition 7.5.3 (Motion Primitive)**.** A motion primitive of a group of joints G is defined by the trajectory segments of all joints $j \in G$ corresponding to a common temporal interval $I = [t_{start}, t_{end}] \subset \mathbb{R}$ such that $P(t_{start}; t_{end}) > P(\rho; t_{start}) \ \forall \rho \in (t_{start}, t_{end})$. Namely

$$\gamma_G^I = \{\xi_{j_1}(t), \ldots, \xi_{j_K}(t)\} \text{ for } t \in [t_{start}, t_{end}]. \tag{7.8}$$

Primitive discovery in an activity A set of primitives is extracted from an entire sequence of an activity ς by sequentially finding the time instances which maximize (7.6).

Let t_0 and t_{seq} denote the starting and ending instances of the sequence, respectively. Let also

$$t_n = \underset{\rho \in [t_{n-1}, t_{seq}]}{\arg\max}\ P(\rho; t_{n-1}), \tag{7.9}$$

and $\mathcal{I}_\varsigma = \{[t_{n-1}, t_n] \,|\, n \in \mathbb{N} \text{ and } t_n \leq t_{seq}\}$ the set of time intervals defining successive motion primitives in the sequence. The set of motion primitives discovered in the entire sequence ς is given by

$$\Gamma_G^\varsigma = \{\gamma_G^I \,|\, I \in \mathcal{I}_\varsigma\}. \tag{7.10}$$

As noted in the introduction, and also shown in Figure 7.5, there is a significant motion variation across subjects, activities and sampling rates. For example, for the upper limbs it is known that the range of motion varies from person to person and is influenced by gait speed (de los Reyes-Guzmán et al., 2014). This is in turn influenced by the specific task, and determining ranges of motion is still a research topic (Gates et al., 2016) (for a review on range of motions for upper limbs, see (de los Reyes-Guzmán et al., 2014)). This makes analysis and recognition of motion primitives taken from different datasets, activities and subjects problematic. To induce invariance with respect to these factors we apply anatomical normalization.

More specifically, the main source of variation of the primitives is due to the anatomical differences among the subjects. To remove the influence of these differences on the primitives we consider a scaling factor k_G based on the length ℓ_G of the limb defined by group G, namely $k_G = 1/\ell_G$. Hence, given a primitive γ_G^I we scale the trajectory of each joint by the constant k_G. By applying the anatomical normalization to the entire collection of motion primitives for group G discovered across all sequences of a dataset $\mathcal{D}$ we obtain the set of motion primitive of the group, namely

$$\Gamma_G = \{\Gamma_G^\varsigma \,|\, \varsigma \in \mathcal{D}\}. \tag{7.11}$$

In Section 7.7 we provide a quantitative evaluation of the normalization effectiveness, together with a comparison with additional normalization candidates.

7.6 Motion Primitive Recognition

In the previous section we have shown that for each group of joints $G_m, m = 1, \ldots, 6$, the motion flux obtains the interval $I = [t_{start}, t_{end}]$ matching the joint trajectory of a sequence in so determining a primitive as a path $\gamma_{G_m}^I : I \subset \mathbb{R} \mapsto \mathbb{R}^9$, given a video sequence of a human activity. Here $\mathbb{R}^9$ is due to the

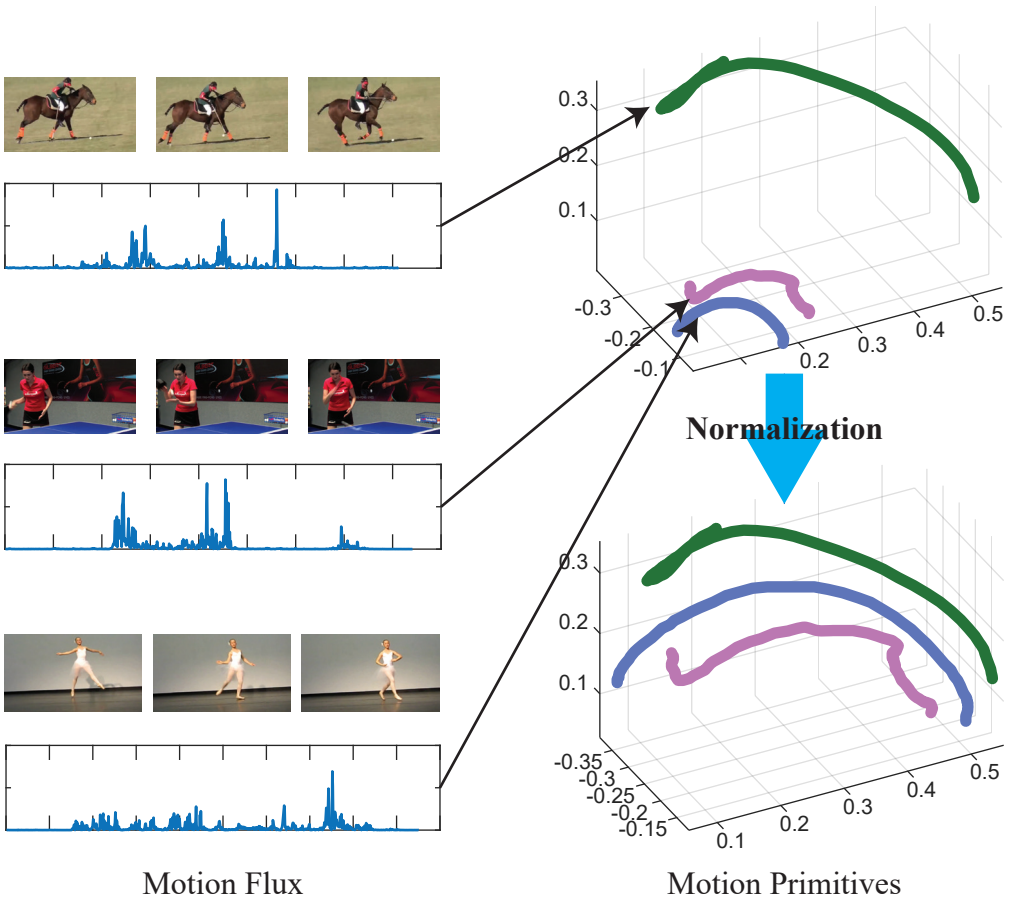
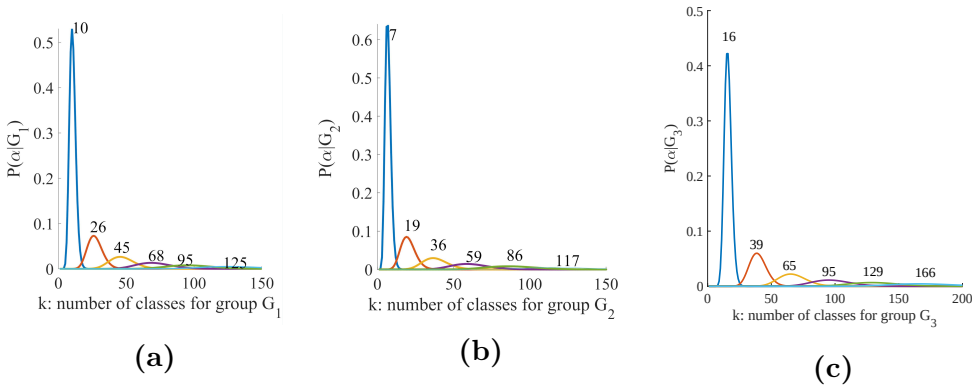


Figure 7.5. Left: Motion flux of three motion primitives of group $G3$ labeled as 'Elbow Flexion', discovered from video sequences taken from the ActivityNet dataset. **Right:** Motion primitives before and after the normalization, for clarity only the curve of the out most joint is shown.(Best seen on screen, zoomed-in)

path being related to the 3 joints of each group G_m, as indicated in Fig. 7.2. We have also seen that the path is normalized by the link length of a limb, to limit variations due to bodies dissimilarities. For clarity from now on we shall denote each primitive with γ unless the context requires to add superscripts and subscripts, and in general subscripts and superscripts are local to this section, also we shall refer to the group a primitive or trajectory belongs to both with G_m and more in general with G.

We expect that the following facts will be true of the discovered motion primitives:

1. Each primitive of motion is independent of the gender, (adult) age, and body structure, under normalization.
2. Each primitive of motion can be characterized independently of the specific activity, hence the same primitive can occur in several activities (see Section 7.7 for a distribution of discovered primitives in a set of



(a) **(b)** **(c)**

Figure 7.6. Number k of components for groups G_1, G_2 and G_3. Values of k are computed adjusting α so as to maximize the posterior $p(\alpha, G_m)$, given the data, namely the sampled primitives in the groups.

activities).

3. The motion flux ensures that each unknown segmented primitive belongs to a class such that: the number of classes is finite and the set of classes can be mapped onto a subset of motion primitives defined in biomechanics (see e.g table 1.1 of (Hamill and Knutzen, 2006)).

To show experimentally the above results we shall introduce a hierarchical classification. The hierarchical classification first partitions the primitives of each group into classes. Once the classes are generated a class representative is chosen and inspected to assign a label to the class. We show that the classes correspond to a significant subset of the motion primitives defined in biomechanics, thus ensuring a proper partition. Each class is then further partitioned into subclasses to comply with the inner diversification of each class of primitives. This last classification is further used for recognition of unknown discovered primitives.

Primitive recognition is used to both test experimentally the three above results of the introduced motion flux method and for applications where discovering and recognition of primitives of human motion is relevant (see for example (Abernethy, 2013)).

7.6.1 Solving primitive classes

We describe in the following the method leading to the generation of all the primitive classes illustrated in Fig 7.12.

We consider three MoCap datasets (Mandery et al., 2015; Ionescu et al., 2014; CMU,) guaranteeing the ground truth for the human pose and segment the activities according to the motion flux method, described in the previous section. Let Γ_G be the set of primitives collected for group G according to equation (7.11). Let $\gamma_\nu \in \Gamma_G$, $\nu = 1, \ldots, S$, with S the number of primitives in Γ_G, $\gamma_\nu = (\xi^\nu_{j_1}, \xi^\nu_{j_2}, \xi^\nu_{j_3})$ is formed by the trajectories of the joints in G. Out

of these trajectories we choose the one of the most external joint (see Figure 7.2) that we indicate with ξ_E^ν. We order these trajectories, each designating a primitive in group G, with an enumeration $\langle \Gamma_G \backslash \xi_E \rangle_{\nu=1}^S$, S the number of discovered primitives for group G. Note that we can arbitrarily enumerate the primitives of a group, restricted to a single joint, though they are unlabeled and unknown, and this is what the first model should solve.

At this step, model generation amounts to find the classes of primitives for each group G, taking the trajectories ξ_E^ν in the enumeration $\langle \Gamma_G \backslash \xi_E \rangle_{\nu=1}^S$ as observations.

Feature Vectors Given a trajectory ξ_E^ν, with ν the index in the enumeration $\langle \Gamma_G \backslash \xi_E \rangle_{\nu=1}^S$, a feature vector is obtained by first computing curvature $\kappa(s(t))$ and torsion $\tau(s(t))$ on the trajectory ξ_E^ν, where $s(t)$ indicates the arc length as already defined in Section 7.5 for trajectories. Then we take three contiguous points $(x_{i-1}, y_{i-1}, z_{i-1}), \ldots, (x_{i+1}, y_{i+1}, z_{i+1})$ on the trajectory $\hat{\xi}_E^\nu$ decimated by a factor of 5 (Alt and Guibas, 2000), keeping the curvature and torsion of the sampled points, after decimation. We choose curvature and torsion as they suffice to specify a 3D curve up to a rigid transformation. The formed feature vector is indicated by $\mathcal{F}_i$, where the index i is the index of the middle point (x_i, y_i, z_i), it is of size 17×1 and it is defined as follows:

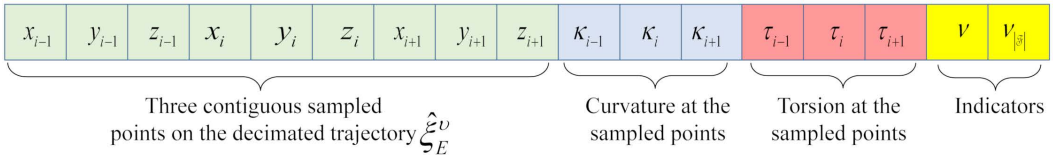


Figure 7.7. Transposed feature vector of 3 contiguous sampled points on the decimated trajectory.

The last two elements $\nu, \nu_{|\mathcal{F}_i|} \in \mathbb{R}$ of $\mathcal{F}_i$ are indicators. Namely, the indicator ν is the index, in the enumeration $\langle \Gamma_G \backslash \xi_E \rangle_{\nu=1}^S$, identifying the trajectory the 3 points belong to, the three points are the first 6 element of the feature vector. On the other hand, the indicator $\nu_{|\mathcal{F}_i|}$ specifies the number of features vectors the decimated trajectory $\hat{\xi}_E^\nu$ is decomposed into, here $|\cdot|$ indicates the cardinality; These two indicators, allow to recover the path a feature vector belongs to, and are normalized and denormalized as follows. Let $\mathbb{F}_G^\xi$ be the set of all feature vectors for the trajectories in $\langle \Gamma_G \backslash \xi_E \rangle_{\nu=1}^S$, and let their number be W. Accordingly, let $\nu_{|\mathcal{F}|} = (\nu_{|\mathcal{F}_1|}, \ldots, \nu_{|\mathcal{F}_W|})$, then the normalization and denormalization for the element $\nu_{|\mathcal{F}_i|}$ (and similarly for ν) is defined as follows, with g indicating the denormalization:

$$\begin{aligned} \hat{\nu}_{|\mathcal{F}_i|} &= \frac{\nu_{|\mathcal{F}_i|} - min(\nu_{|\mathcal{F}|})}{max(\nu_{|\mathcal{F}|}) - min(\nu_{|\mathcal{F}|})} \\ g(\hat{\nu}_{|\mathcal{F}_i|}) &= \hat{\nu}_{|\mathcal{F}_i|}(max(\nu_{|\mathcal{F}|}) - min(\nu_{|\mathcal{F}|})) + min(\nu_{|\mathcal{F}|}) \end{aligned} \tag{7.12}$$

Generation of the primitives classes Given the feature vectors for each trajectory in the enumeration $\langle \Gamma_G \backslash \xi_E \rangle_{\nu=1}^{S}$, the goal is to cluster them and return a cluster for each class of primitives. Since we do not even know the number of classes the primitives should be partitioned into, a good generative model to approximate the distribution of the observations is the Dirichlet process mixture (DPM) (Ferguson, 1973; Antoniak, 1974). The Dirichlet process assigns probability measures to the set of measurable partitions of the data space. This induces in the limit a finite mixture since, by the discreteness of the distributions sampled from the process, parameters have positive probability to take the same value, in so realizing components of the mixture. Here we assume that feature vectors in the data space are realizations of normal distributions with a conjugate prior. Namely the variables have precision priors following the Wishart distribution and location parameters prior following the normal distribution. The Dirichlet mixture model is based on the definition of a Dirichlet process $\Pi(\cdot, \cdot)$ with $\Pi \sim DP(H, \alpha)$ (D being the Dirichlet distribution), where H is the base distribution and α the precision parameter of the process (see (Teh, 2011)). In the Dirichlet process mixture the value of the precision α of the underlying Dirichlet process influences the number of classes generated by the model.

For determining the number of classes for each group G we estimate the posterior $P(\alpha|G)$, of the precision parameter α according to a mixture of two gamma distributions, as described in (West, 1992), choosing the best value. This is a rather complex simulation process since it requires different initializations of the parameters of the gamma distribution for α within the estimation of the parameters of the DPM, for each group G. Here the parameters of the DPM are estimated according to (Jain and Neal, 2004). Distributions of α for the groups G_1, G_2 and G_3, according to different simulation processes, are given in Fig. 7.6 where the number of components k for the maximum values of each distribution, are indicated. Finally the DPM returns the parameters of the components (for each group G) given the feature vector $\mathcal{F}_i$, as:

$$\begin{aligned} &\Theta_G = \langle k, \{\Theta_w \mid \Theta_w = (\pi_w, \mu_w, \Sigma_w), w = 1, \ldots, k\}\rangle,\ k \geq 1. \\ &p(\mathcal{F}_i|\Theta_G) = \textstyle\sum_{w=1}^{k} \pi_w \mathcal{N}(\mathcal{F}_i|\mu_w, \Sigma_w). \end{aligned} \tag{7.13}$$

Note that the number of components k is unknown and estimated by the DPM, hence it is one of the parameters for each group. The parameters μ_w and Σ_w are the mean vector and covariance matrix of the w-th Gaussian component of the mixture, indicated by $\mathcal{N}$, and π_w is the w-th weight of the mixture, with $\sum_w \pi_w = 1$. Hence, $p(\mathcal{F}_i|\Theta_G)$ is the probability of the feature vector $\mathcal{F}_i$, given the parameters Θ_G.

We expect that each $\Theta_w \in \Theta_G$ indicates the parameters of a component C_w^G, collecting primitives of the same type, in group G. In other words, we expect that two feature vectors, say $\mathcal{F}_p, \mathcal{F}_q$, of group G, belong to the same component if their likelihood are both maximized by the same parameters $\Theta_w \in \Theta_G$.

Assigning primitives to classes The classification returns, for each group G_m, the number k of components indicated in Fig. 7.12, say $k = 10$ for G_1, G_5, G_6, $k = 7$ for G_2 and $k = 16$ for G_3, G_4, also thanks to the specification of the α parameter, as highlighted above (see Fig. 7.6). Components are formed by features vectors. To retrieve the trajectories and generate a corresponding class of primitives, ready to be labeled, we use the normalized indicators placed in position 16th and 17th of the feature vector (Fig 7.7) and the denormalization function g. Let $C_w^{G_m}$ be a component of the mixture of the group G_m, identified by parameters $\Theta_w \in \Theta_{G_m}$. Algorithm 5 shows how to compute the class of primitives:

Algorithm 5: Obtaining classes of primitives from DPM components. Here $|\cdot|$ indicates cardinality.

Input: Component $C_w^{G_m}$ of DPM
Output: Class $\mathcal{L}_w^{G_m}$ of primitives
Initialize $U_{\xi_E}^{\nu} = \emptyset$, $\nu = 1, \ldots, S$, S number of primitives in Γ_{G_m}
foreach *Feature vector* $\mathcal{F}_i$ *in* $C_w^{G_m}$ **do**
    compute $g(\nu)$ and associate it with the trajectory ξ_E^{ν};
    $U_{\xi_E}^{\nu} = \{\mathcal{F}_i\} \cup U_{\xi_E}^{\nu}$;
    compute $g(\nu_{|\mathcal{F}|})$, number of feature vectors the trajectory ξ_E^{ν} is decomposed into;
if $|U_{\xi_E}^{\nu}| \geq 0.8 g(\nu_{|\mathcal{F}|})$ **then**
    find the primitive $\gamma_\nu \in \Gamma_{G_m}$ designated by ξ_E^{ν}
    assign the pair (γ_ν, Θ_w) to $\mathcal{L}_w^{G_m}$
return Class $\mathcal{L}_w^{G_m}$.

At this point we have generated the classes $\mathcal{L}_w^{G_m}$, $w = 1, \ldots, k$, $k \in \{7, 10, 16\}$ of primitive for each group G_m. To label the classes we proceed as follows. Let $p(\gamma_\nu|\Theta_w) = 1/g(\nu_{|\mathcal{F}|}) \sum_i p(\mathcal{F}_i|\Theta_w)\delta(\mathcal{F}_i)$, where $\delta(\mathcal{F}_i) = 1$ if $\mathcal{F}_i \in U_{\xi_E}^{\nu}$ and 0 otherwise. For each class $\mathcal{L}_w^{G_m}$ the class representative is the primitive maximizing $p(\gamma_\nu|\Theta_w)$. The representative primitive is observed and labeled by inspection, according to the nomenclature given in biomechanics, see (Hamill and Knutzen, 2006). The same label is assigned to the class $\mathcal{L}_w^{G_m}$, without need to inspect all other primitives assigned to the class.

Average Hausdorff distances between each primitive in a class and its class representative, for each class in group G_2, are given in Table 7.1, where classes for G_2 are enumerated according to the labels illustrated in Fig. 7.12. Note that in Table 7.1 R_w is the class representative, so $R_w \in \mathcal{L}_w^{G_m}$, $w = 1, \ldots, 7$; $\forall \xi_{E \setminus R_w}$ abbreviates $\forall \xi_E \in \mathcal{L}_w^{G_2}, \xi_E \neq R_w$. Finally, $\mathcal{L}_w$ abbreviates $\mathcal{L}_w^{G_2}$. Note that distances with elements of other classes are obviously not considered, hence the dashes in other classes columns.

Table 7.1. Average Hausdorff distance to each class representative in G_2

	R_1	R_2	R_3	R_4	R_5	R_6	R_7
$\forall\xi_{E\setminus R_1}\in\mathcal{L}_1$	0.121	-	-	-	-	-	-
$\forall\xi_{E\setminus R_2}\in\mathcal{L}_2$	-	0.173	-	-	-	-	-
$\forall\xi_{E\setminus R_3}\in\mathcal{L}_3$	-	-	0.144	-	-	-	-
$\forall\xi_{E\setminus R_4}\in\mathcal{L}_4$	-	-	-	0.112	-	-	-
$\forall\xi_{E\setminus R_5}\in\mathcal{L}_5$	-	-	-	-	0.081	-	-
$\forall\xi_{E\setminus R_6}\in\mathcal{L}_6$	-	-	-	-	-	0.142	-
$\forall\xi_{E\setminus R_7}\in\mathcal{L}_7$	-	-	-	-	-	-	0.114

7.6.2 Models for recognition

The recognition problem is stated as follows. Given an unlabeled primitive γ_u, for group G_m obtained by segmenting an activity (from any dataset) with the motion flux method, γ_u is labeled by the label of class $\mathcal{L}_w^{G_m}$, if:

$$p(\gamma_u|\Theta_w) > p(\gamma_u|\Theta_i), \quad \forall i, i \neq w \tag{7.14}$$

We found experimentally that relying on the same parameters used for finding the classes of primitives, described in the previous sub-section, does not lead to optimal results. In fact, recomputing a DPM model for each class and introducing a loss function on the set of hypotheses, computed by thresholding the best classes, leads to an improvement up to the 20% in the recognition of an unknown primitive.

To this end we compute a DPM for each class $\mathcal{L}_w^{G_m}$ using as observations the primitives collected in the class, by Algorithm 5. Therefore the generated DPM model $\mathcal{M}_w$ for each class $\mathcal{L}_w^{G_m}$ is made by a number of components with parameters $\Theta_w = \{\Theta_{w_1}, \ldots, \Theta_{w_\rho}\}$, with ρ varying according to the components generated for class $\mathcal{L}_w^{G_m}$. The number of components mirrors the idiosyncratic behavior of each class of primitives, therefore ρ varies for each class $\mathcal{L}_w^{G_m}$. To generate these DPM models we use all the three trajectories of the primitives $\gamma \in \mathcal{L}_w^{G_m}$, and for each of them we use the same decimation and feature vector as shown in Fig. 7.7.

Given the refined classification, the recognition problem, at this point, is stated as follows. Let $\gamma_u = (\xi_{u_1}, \xi_{u_2}, \xi_{u_3})$ be an unknown primitive, of a specific group G, and let $\{\mathcal{F}_{u_1}, \ldots, \mathcal{F}_{u_q}\}$ be the set of features the three trajectories are decomposed into. Then $\gamma_u \in \mathcal{L}_w^{G_m}$, hence is labeled by the label of this class, if:

$$p(\mathcal{F}_{u_1}, \ldots, \mathcal{F}_{u_q}|\Theta_w) = \sum_{j=1}^{\rho} \pi_j \prod_{n=1}^{q} p(\mathcal{F}_{u_n}|\Theta_{w_j}) > p(\mathcal{F}_{u_1}, \ldots, \mathcal{F}_{u_q}|\Theta_h) = \sum_{j=1}^{\rho'} \pi'_j \prod_{n=1}^{q} p(\mathcal{F}_{u_n}|\Theta_{h_j}) \tag{7.15}$$

for any parameter set Θ_h associated with a class $\mathcal{L}_h^{G_m}$ of the group G_m. Here π_j and π'_j are the mixture weights, with $\sum_j \pi_j = 1$ and ρ, ρ' indicate the number

of components of the chosen models. For example, the model of class $\mathcal{L}_w^{G_2}$, with $w = 1$, will have a set of parameters $\Theta_w = \{\Theta_{w_1}, \ldots, \Theta_{w_\rho}\}$, while the model of class $\mathcal{L}_{w'}^{G_2}$, with $w' = 3$, will have a set of parameters $\Theta_{w'} = \{\Theta_{w'_1}, \ldots, \Theta_{w'_{\rho'}}\}$, with $w_\rho \neq w'_{\rho'}$.

This formulation is much more flexible than (7.14), also because it computes the class label by considering all the components and therefore it does not care whether the features are scattered amid components, and does not need to reconstruct the whole trajectories as was done for generating the classes of primitives. Furthermore, under this refined classification we can improve (7.15) considering a geometric measure to reinforce the statistics measure in the choice of the class label for γ_u.

More precisely, let us form a set of hypotheses for an unknown primitive with feature set $\{\mathcal{F}_{u_1}, \ldots, \mathcal{F}_{u_q}\}$ as follows (we are still assuming a specific group G_m):

$$\mathbb{H} = \{\langle C_{w_j}, \Theta_{w_j}\rangle | \prod_{n=1}^{q} p(\mathcal{F}_{u_n}|\Theta_{w_j}) > \eta, \langle C_{w_j}, \Theta_{w_j}\rangle \in \mathcal{M}_w, w = 1, \ldots, k\} \quad (7.16)$$

Namely C_{w_j} is a component of the DPM $\mathcal{M}_w$, with $w = 1, \ldots, k$, k the number of classes in group G_m, and $j = 1, \ldots, \rho$, such that the associated parameter Θ_{w_j} makes the joint probability of the features, the primitive is decomposed into, greater than a threshold η. This means that we are collecting in $\mathbb{H}$ those components coming from all the models of group G_m, whose joint probability of the feature set of the unknown primitives γ_u forms an hypotheses set, or a set from which we can select the correct label to assign to γ_u.

The advantage of the hypotheses set is that we delay the decision of choosing the labeled class for the unknown primitive to further evidence, which we collect by using geometric measures. The role of these geometric measures is essentially to evaluate the similarity between the curve segments coming out from the features of γ_u and those coming from the observations which are indexed in the components in $\mathbb{H}$. In the following we succinctly describe the new geometric features, which are computed as follows, both for the features of the unknown primitive γ_u and for the features coming from the observations indexed in C_{w_j}. Let us consider any pair $\langle C_{w_j}, \Theta_{w_j}\rangle \in \mathbb{H}$, by definition (7.16), C_{w_j} indexes features $\{\mathcal{F}_{\nu_1}, \ldots, \mathcal{F}_{\nu_s}\}$, s varying according to the specific component C_{w_j}. For each of these features we consider the points of the trajectory ξ^ν, recovered from the decimated trajectory $\hat{\xi}_\nu$, between $(x_{i-1}, y_{i-1}, z_{i-1})$ and $(x_{i+1}, y_{i+1}, z_{i+1})$. Let us consider these curve segments, which we combine whenever they occur in sequence in C_{w_j} and call any of these curve segments $\mathbf{y}$. In particular, the collection of these segments in C_{w_j} is called the manifold of C_{w_j}, denoted $man(C_{w_j})$, and the collection of segments generated from the features of γ_u is denoted $man(\gamma_u)$, examples are given in Fig. 7.8.

We compute for each $\mathbf{y}$ both in $man(C_{w_j})$ and in $man(\gamma_u)$ the tangent $\mathbf{t}$, normal $\mathbf{n}$ and binormal $\mathbf{b}$ vectors. Based on these vectors, we compute the ruled

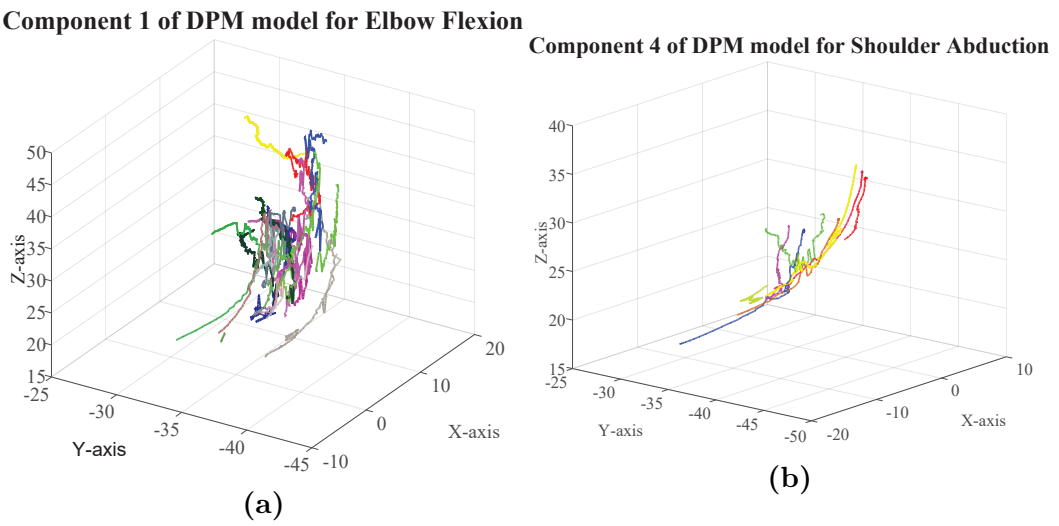


Figure 7.8. Manifold generated by a component of the DPM model for Elbow Flexion on the left and from a component of Shoulder Abduction on the right.

surface $\mathcal{R} = \frac{\mathbf{n}\times\mathbf{n}'}{\|\mathbf{n}\times\mathbf{n}'\|}$, where $\mathbf{n}'$ is the derivative of $\mathbf{n}$. The ruled surface forms a ribbon of tangent planes to the curve segment $\mathbf{y}$. In particular, let us distinguish the curve segments in $man(\gamma_u)$ denoting them $\mathbf{y}_u$. We compute the distances between any curve segment $\mathbf{y} \in man(C_{w_j})$ and $\mathbf{y}_u \in man(\gamma_u)$ as the distance between the projection $\mathbf{y}_\pi$ of $\mathbf{y}$ on the ruled surface tangent to $\mathbf{y}$, and the *closest point* $\mathbf{q}$ of $\mathbf{y}_u$ to $\mathbf{y}_\pi$. We denote this distance $\delta(\mathbf{y}_u, \mathbf{y})$. We consider also the distance between the Frenet frames at closest points $\mathbf{q}$ of $\mathbf{y}_u$ and point $\mathbf{q}'$ of $\mathbf{y}_\pi$ denoted F_R and computed as follows: $F_R(\mathbf{q}, \mathbf{q}') = \text{trace}((\mathcal{I} - R_{\mathbf{q},\mathbf{q}'})(\mathcal{I} - R_{\mathbf{q},\mathbf{q}'})^\top)$, with $\mathcal{I}$ the identity matrix and $R_{\mathbf{q},\mathbf{q}'}$ the rotation, in the direction from $\mathbf{q}$ to $\mathbf{q}'$. Then the cost of a component C_{w_j} in $\mathbb{H}$, given an unknown primitive γ_u, with feature set $\{\mathcal{F}_{u_1}, \ldots, \mathcal{F}_{u_q}\}$, is defined as:

$$Cost(C_{w_j} \in \mathbb{H}|\gamma_u) = max\{\delta(\mathbf{y}_u, \mathbf{y}) + F_R(\mathbf{q}, \mathbf{q}')|\mathbf{y_u} \in man(\gamma_u) \text{ and } \mathbf{y} \in man(C_{w_j})\} \tag{7.17}$$

Note that both $\delta(\mathbf{y}_u, \mathbf{y})$ and $F_R(\mathbf{q}, \mathbf{q}')$ were both computed looking at the minimum distance between a considered curve segment and the projection on the ruled surface of the other curve segment. Hence the component minimizing the above cost and maximizing the probability in (7.15) will indicate the class label, since its related parameter indicates exactly a component of one of the classes $\mathcal{L}_w^{G_m}$. Note that if in (7.15) η is taken to be equal to $max(\prod_{n=1}^{q} p(\mathcal{F}_{u_n}|\Theta_{w_j}))$ then $\mathbb{H}$ would have only a single element $\langle C_{w_j}, \Theta_{w_j}\rangle$. Hence to find the correct label for γ_u we push η as high as possible using the above cost. More precisely, the component of the class $\mathcal{L}_w^{G_m}$ which should label the unknown primitive γ_u is computed as follows:

$$C^\star = \arg\min_{C_{w_j}} \sup_{\eta} \{Cost(C_{w_j} \in \mathbb{H}|\gamma_u)| \prod_{n=1}^{q} p(\mathcal{F}_{u_n}|\Theta_{w_j}) > \eta\} \tag{7.18}$$

To conclude this section we can note that the computation of the hierarchical model that first generates the primitive classes and then uses these generated sets to estimate model parameters to be used in the recognition of an unknown primitive, has an exponential cost, in the dimension of the features and in the size of the observations. However using the computed models to recognize an unknown primitive is $\mathcal{O}(n^2 log\ n)$ where n is the size of γ_u, since all the curve segments in the models can be precomputed together with the models. Results on both the primitive generation and on recognition are given in the next section.

7.7 Experiments

In this section we evaluate the proposed framework for discover and classification of human motion primitives. For all the evaluations we consider three reference MoCap public datasets (Mandery et al., 2015; Ionescu et al., 2014; CMU,).

First we evaluate the accuracy of the motion primitives discovered using the motion flux, further we evaluate the accuracy of the classification and recognition. Additionally, we examine the distribution of recognized primitives with respect to the type of performed activity on the ActivityNet dataset (Ghanem et al., 2017). Finally, we address the dataset of human motion primitives we have created, which consists of the primitives discovered on the three reference MoCap datasets using the motion flux, and the DPM models established for each primitive category.

7.7.1 Reference Datasets

The datasets we consider for the evaluation of the motion flux are the Human3.6M dataset (H3.6M) (Ionescu et al., 2014), the CMU Graphics Lab MoCap database (CMU) (CMU,) and the KIT Whole-Body Human Motion Database (KIT-WB) (Mandery et al., 2015). The sampling rates used in these datasets are 50Hz for H3.6M, 60/120Hz for CMU and 100Hz for KIT-WB. In order to have the same sampling rate for all sequences we have transformed all of them to 50Hz. The pose of the joints specified in Fig. 7.2 are extracted for each frame of the sequences as described in the preliminaries, considering the ground-truth 3D poses. For KIT-WB the trajectories of the joints are computed from the marker positions taken from the C3D files. We considered 40 activities from the three reference datasets. Fig. 7.9 shows the total number of motion primitives discovered for the five most general activities according to the ActivityNet taxonomy based on the motion flux for each group G_m. Table 7.2 shows the total number of motion primitives discovered from the three datasets.

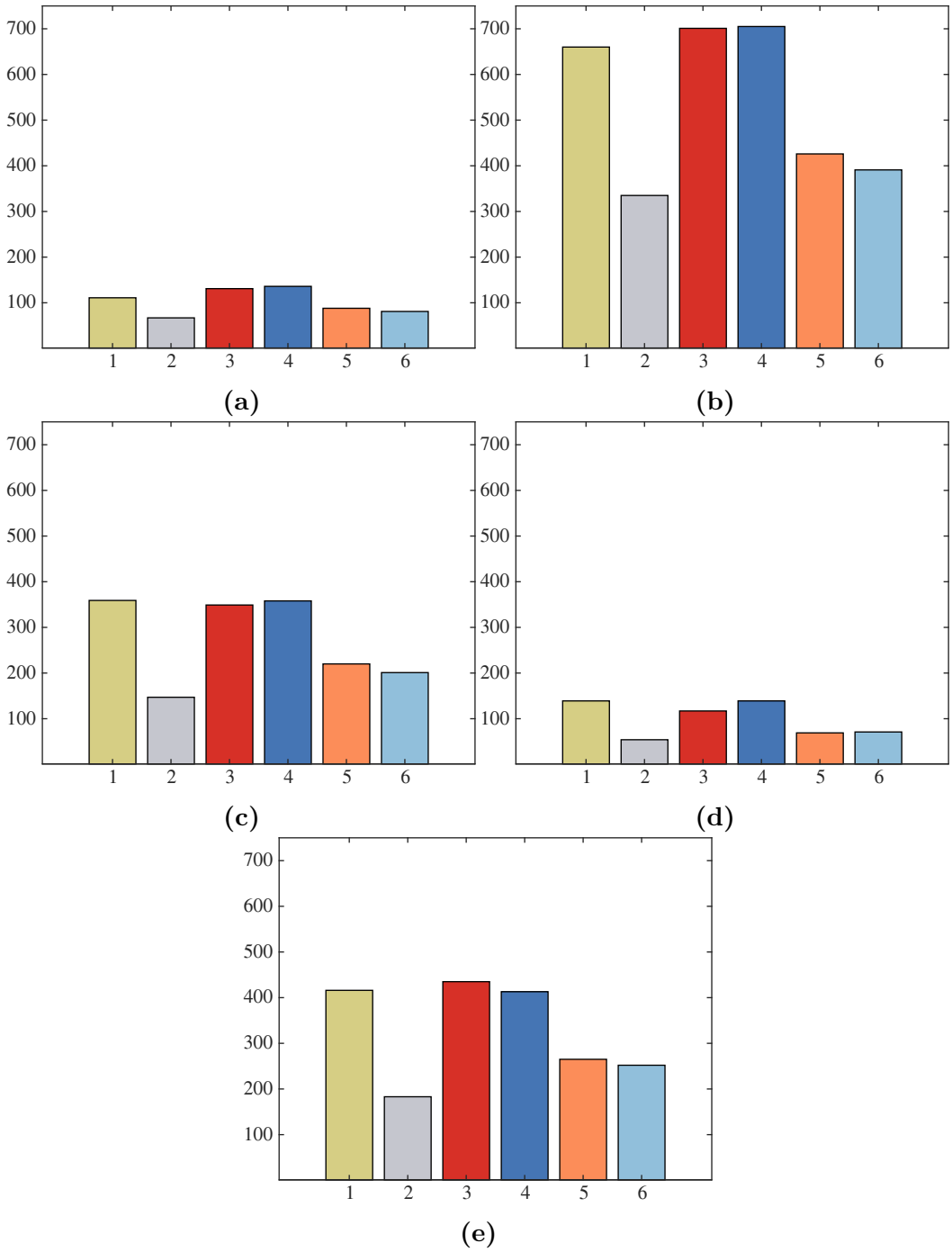


Figure 7.9. Total number of discovered primitives for each group for the five most general categories of the ActivityNet dataset. Clock-wise from top-left: *Eating and drinking Activities*; *Sports, Exercise, and Recreation*; *Socializing, Relaxing, and Leisure*; *Personal Care*; *Household Activities*. Each color corresponds to a different group following the convention of Fig. 7.12. Note: Axes scale is shared among the plots.

Table 7.2. Total number of unlabeled primitives discovered for each group using the motion flux on the reference datasets

	G1	G2	G3	G4	G5	G6
Total	1665	759	1773	1703	1152	1015

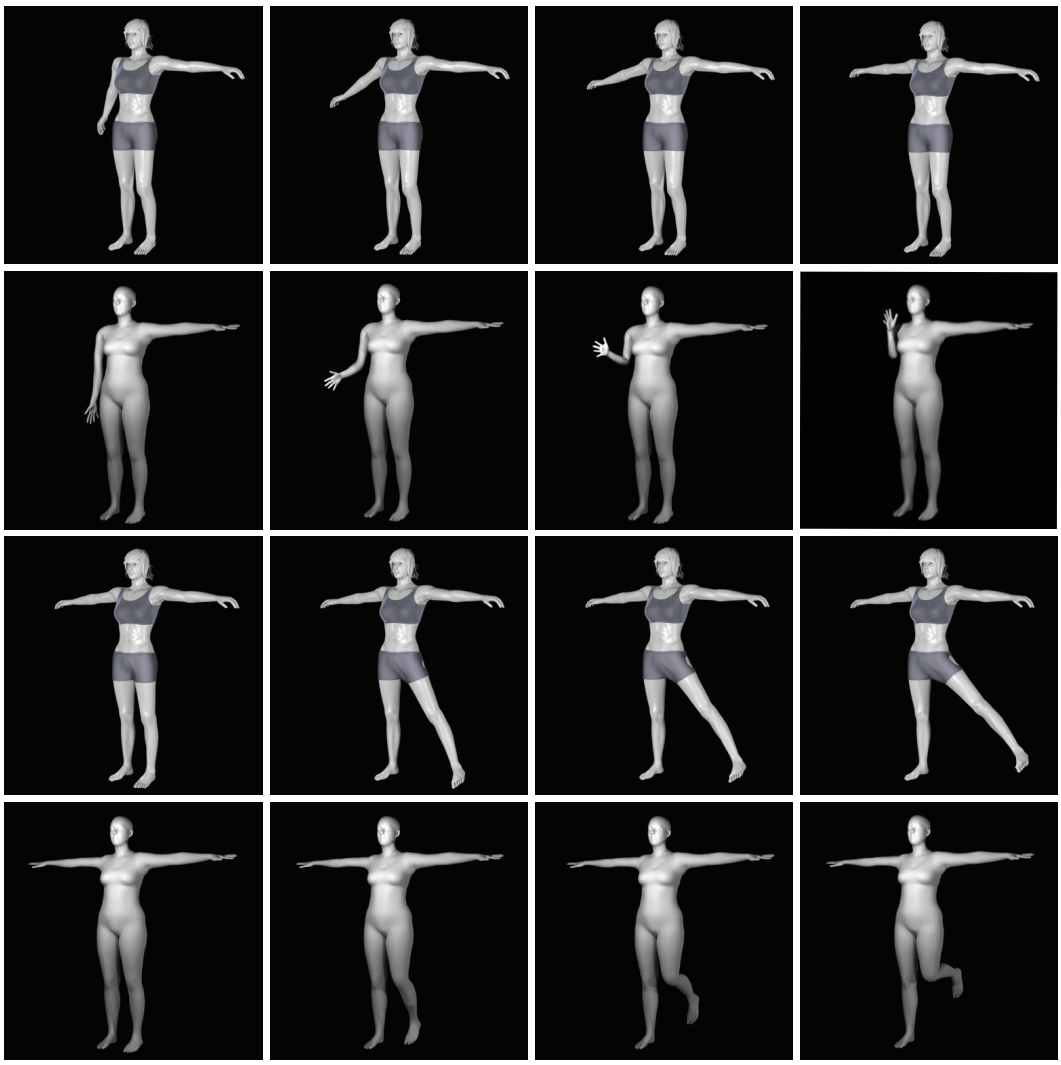


Figure 7.10. Example of synthetic motion primitive, specifically right arm Shoulder Abduction (first row) and Elbow Flexion (second row), left leg Hip abduction (third row) and Knee Flexion (fourth row). For each synthetic motion primitive the four imaged poses match four representative poses extracted from the animation of the aforementioned primitive.

7.7.2 Motion Primitive Discovery

To evaluate the accuracy of primitive discovery based on the motion flux, we created a baseline relying on a synthetic dataset of motion primitives. This

was necessary to mitigate the difficulty in measuring accuracy, due to the lack of a ground truth.

The synthetic dataset of motion primitives we created is formed by animations of 3D human models for each of the 69 primitive classes discovered in Sec. 7.6. The human models were downloaded from the dataset provided by (Loper et al., 2015) or acquired from (tur, ; ren,). To obtain further characters the shapes of the human models were randomly modified taking care of human height and limb length limits.

Animations of the characters were produced moving the skeleton joints belonging to the 3D human models from a start pose to an end pose representing the primitives. Specifically, for each primitive of each skeleton group the animation was generated in Maya or Blender (depending on the 3D human model format) moving the group joints according to angles, gait speed and limbs proportions as described in (de los Reyes-Guzmán et al., 2014; Gates et al., 2016; Hamill and Knutzen, 2006; Abernethy, 2013).

The dataset reference skeleton, see Fig. 7.2 is matched with the 3D human mesh models by fitting the joint poses of the synthetic data to the reference skeleton, basing on MoSh (Loper et al., 2014; Varol et al., 2017). Examples of synthetic motion primitives, namely the primitives Shoulder abduction and Elbow flexion for the right arm, and Hip abduction and Knee flexion for the left leg, are illustrated in Fig. 7.10, where for each primitive four representative poses extracted from the animations are shown.

The baseline for evaluating accuracy was created generating 4500 random length sequences of synthetic motion primitives placing them one after another in a random order. Between two consecutive primitives a transition phase from the end pose of the preceding one to the beginning pose of the subsequent one was added.

With this procedure we know precisely the endpoints of each primitive.

Then we applied the 'motion flux' method described in Sec. 7.5 to the 3D joints trajectories extracted from the automatically generated sequences and collected the end points of the discovered primitives.

We use the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics to assess the accuracy of the collected endpoints with respect to the known end points in the generated sequences. Let S be the total number of generated sequences. Let $\{\hat{e}_{i,s}\}_{i=1}^{N_G^{(s)}}$ be the i-th automatically discovered endpoint based on the motion flux for the generated sequence $s = \{1, \ldots, S\}$, with $N_G^{(s)}$ the number of primitives for Group G and sequence s. Denoting $\{\bar{e}_{i,s}\}_{i=1}^{N_G^{(s)}}$ the i-th endpoint in the generated sequence s, the MAE and RMSE metrics are defined as follows:

$$MAE = \frac{1}{S}\sum_{s=1}^{S}\frac{\sum_{i=1}^{N_G^{(s)}}|\bar{e}_{i,s}-\hat{e}_{i,s}|}{N_G^{(s)}}, \quad RMSE = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\frac{\sum_{i=1}^{N_G^{(s)}}(\bar{e}_{i,s}-\hat{e}_{i,s})^2}{N_G^{(s)}}}.$$

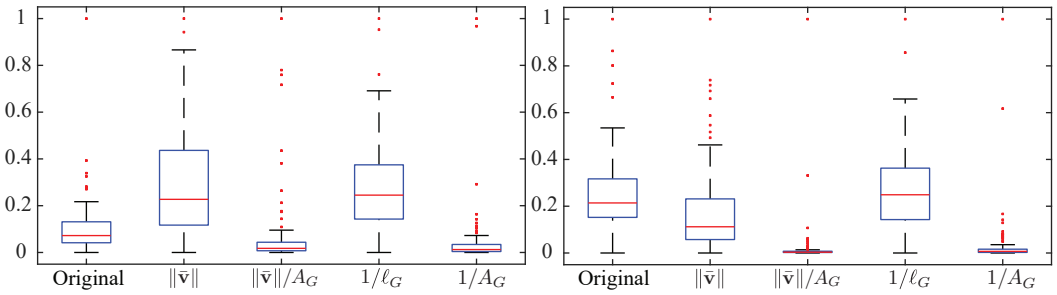


Figure 7.11. Arc length distribution of original and scaled primitives of a specific category for group G_1 (**left**) and G_4 (**right**). The first box in each box plot, corresponds to the original arc length distribution, the next four are the arc length distributions obtained scaling the primitives original data using the detailed scaling factors. Each box indicates the inner 50th percentile of the trajectory data, top and bottom of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, crosses are the outliers.

Results shown in Table 7.3 prove that the proposed method discovers motion primitives quite accurately, since the endpoints are close to those of the automatically generated sequences.

Table 7.3. Accuracy of discovered primitive endpoints (in number of frames)

	G1	**G2**	**G3**	**G4**	**G5**	**G6**	**Overall**
MAE	2.8	3.2	2.9	3.4	3.6	4.1	3.3
RMSE	3.7	4.2	4.1	4.6	4.8	5.2	4.4

Furthermore, to evaluate the effects of the normalization in Fig. 7.11 we show the arc length distribution of motion primitives with and without normalization, as well as considering different normalization constants.

For comparison we consider alternative normalization constants based on anatomical properties and execution style. Specifically, we consider normalization based on the average velocity along $\gamma \in \Gamma_G$, denoted as $\|\bar{\mathbf{v}}\|$, and based on the area A_G covered by group G during its motion. The first is related to the execution speed of the motion and the sampling rate of the data, while the latter is considering anatomical differences among the subjects.

In Fig. 7.11 the first box in each plot corresponds to the original distribution and the following boxes correspond to the distributions resulting by scaling the original one with $\|\bar{\mathbf{v}}\|$, $\|\bar{\mathbf{v}}\|/A_G$, $1/\ell_G$, and $1/A_G$, respectively. We note that normalizing the primitives based on the inverse of the limb length, i.e. ℓ_G, consistently results to an arc length distribution closer to the normal, minimizing the number of outliers indicated by red crosses in the figure. This result is consistent across different activities and groups justifying the choice

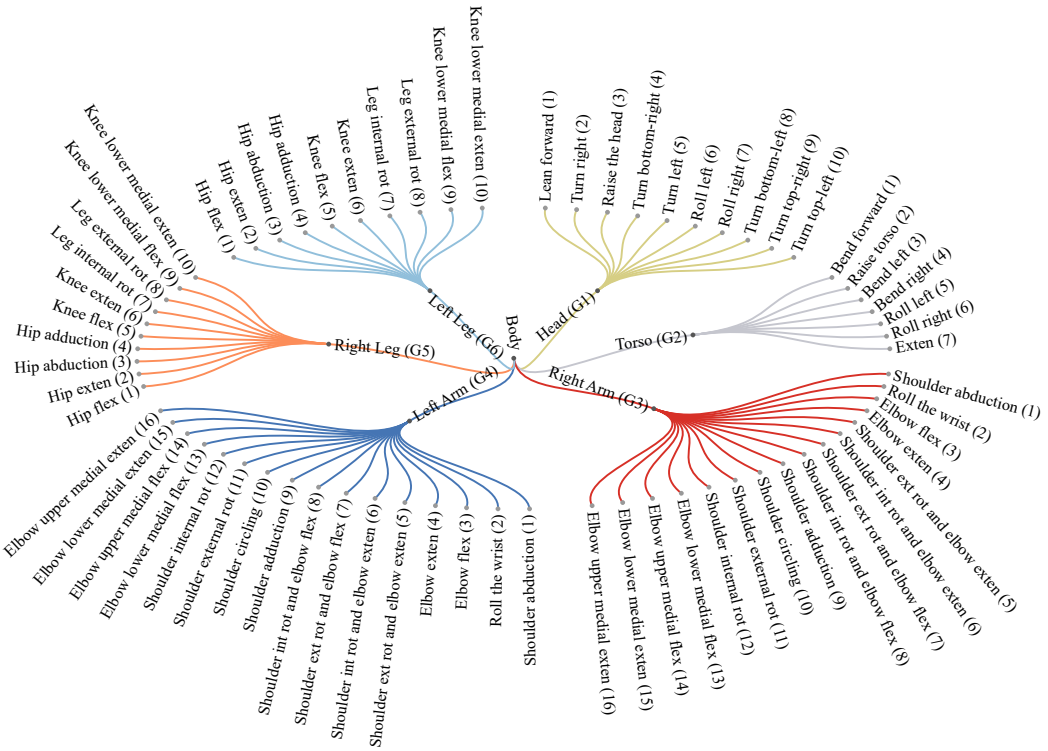


Figure 7.12. Diagram showing the motion primitives of each group. Abbreviation *ext* stands for external, *int* for internal, *rot* for rotation, *exten* for extension, and *flex* for flexion.

of $k_G = 1/\ell_G$ for anatomical normalization.

7.7.3 Motion Primitive Classification and Recognition

As discussed in Section 7.6, the set of primitive categories for each group is generated by a DPM model given the collection of discovered primitives as observations. In this way a total of 69 types of primitives were identified, each described by the distribution parameters. By inspecting a representative primitives for each category, we observed that they correspond to a subset of motion primitives defined in biomechanics. Therefore we generated new DPM models to obtain parameters and corresponding labels for each category. The labeled collection of motion primitives is depicted in Fig. 7.12.

To evaluate the coherence of the generated classes we performed 10 cycles of random sampling, with a rate of 10% at each cycle, of the primitives in each class and verified the class consistency. Only $\sim 2\%$ of the primitives were not correctly classified, according to the label assigned to the class.

For the recognition we adopted the protocol P2 used for pose estimation (see (Sanzari et al., 2016; Tekin et al., 2016)) using one specific subject for testing. Table 7.4 presents the average accuracy of the recognition for each

group, as well as an ablation study with respect to the components of the cost function used in eq. (7.18). Fig. 7.13 shows the corresponding confusion matrices. The results suggest that the DPM classification together with the proposed recognition method capture the main characteristics of each motion primitive category.

Finally, we evaluate the recognition accuracy by considering the same sequences though computing the subject's pose directly from the video frames using (Sanzari et al., 2016). The corresponding results are shown in parentheses in the last column of Table 7.4. We note that the recognition accuracy decreases in average just by 4% by using the estimated pose.

Table 7.4. Primitive recognition accuracy and ablation study

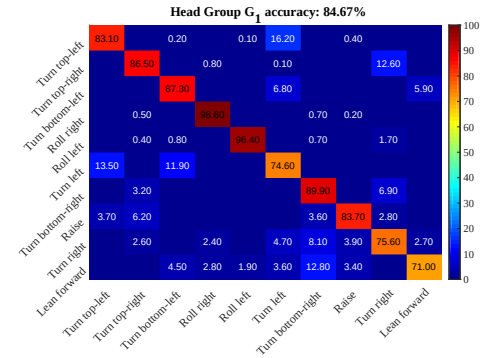
Group	Projection on tangent plane	Frenet frame rotation	Torsion	Curvature	All
G1	0.82	0.80	0.70	0.72	0.84 (0.82)
G2	0.85	0.82	0.75	0.75	0.86 (0.84)
G3	0.80	0.80	0.73	0.74	0.82 (0.78)
G4	0.80	0.79	0.75	0.77	0.83 (0.76)
G5	0.87	0.86	0.72	0.72	0.88 (0.81)
G6	0.86	0.86	0.71	0.73	0.88 (0.82)
Average	0.83	0.82	0.73	0.76	0.85 (0.81)

7.7.4 Primitives in Activities

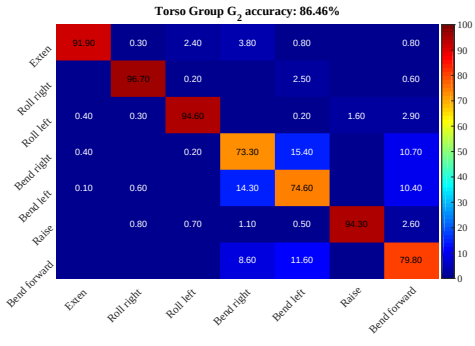
We examine the distribution of discovered motion primitives with respect to the activities been performed by the subjects. We perform our analysis on the sequences of the ActivityNet dataset. More specifically we use the 3D pose estimation algorithm of (Sanzari et al., 2016) on the video sequences of ActivityNet. We then extract motion primitives using the motion flux and perform recognition based on the extracted poses. We consider only the segments of the videos labeled with a corresponding activity. Additionally, we use only the segments were a single subject is detected and at least the upper body is visible. Fig. 7.14 display the distribution of the motion primitives for the five most general activities according to the ActivityNet taxonomy.

7.7.5 Motion Primitives Dataset

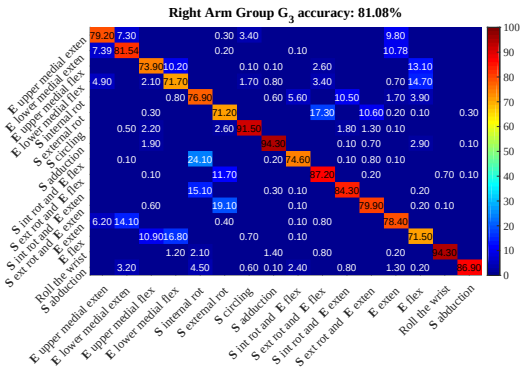
The dataset of annotated motion primitives extracted from the MoCap sequences of H3.6M (Ionescu et al., 2014), CMU (CMU,) and KIT-WB (Mandery et al., 2015) has been made publicly available at `https://github.com/MotionPrimitives/MotionPrimitives`. The dataset provides the start and end frames of each motion primitive together with the corresponding label as well as a reference to the MoCap sequence from which the motion primitive has been extracted.



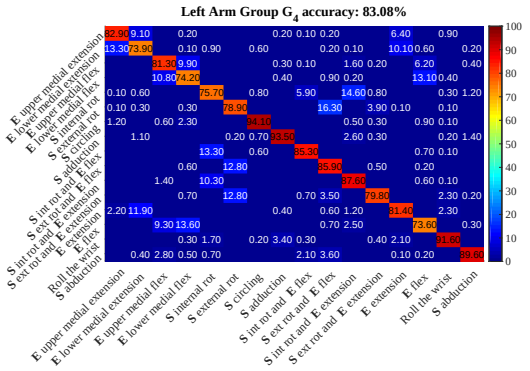
(a)



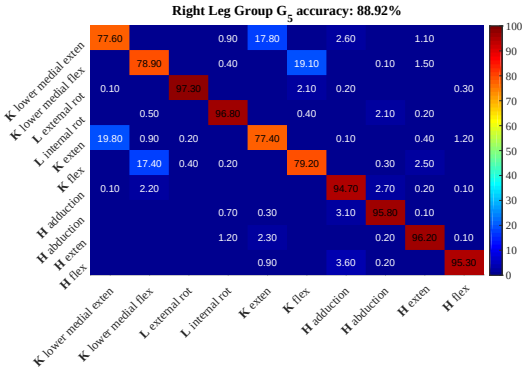
(b)



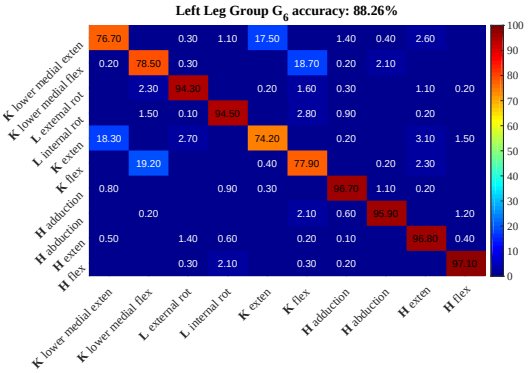
(c) **S** indicates the *shoulder*, **E** the *elbow*.



(d) **S** indicates the *shoulder*, **E** the *elbow*.



(e) **K** indicates the *knee*, **L** the *leg* and **H** the *hip*



(f) **K** indicates the *knee*, **L** the *leg* and **H** the *hip*

Figure 7.13. Confusion matrices for motion primitive recognition. The matrices for G1 and G2 are shown at the top, G3 and G4 at the middle, while G5 and G6 are shown at the bottom.

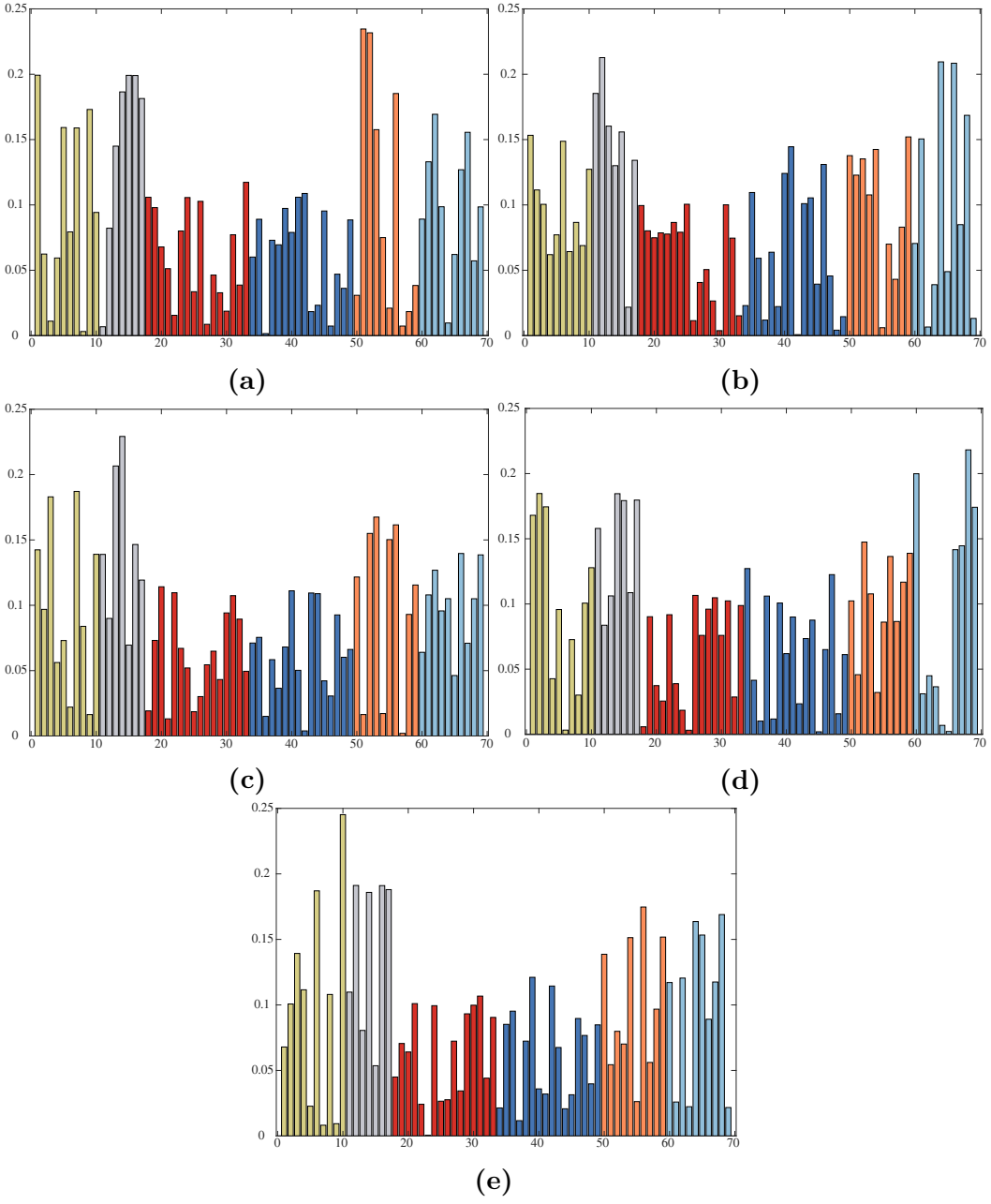


Figure 7.14. Distribution of the 69 primitives for the five most general categories of the ActivityNet dataset. Clock-wise from top-left: *Eating and drinking Activities*; *Sports, Exercise, and Recreation*; *Socializing, Relaxing, and Leisure*; *Personal Care*; *Household Activities*. Each color corresponds to a different group following the convention of Fig. 7.12.

7.7.6 Comparisons with state of the art on motion primitive recognition

We consider here the results of (Holte et al., 2010), so far the only work providing quantitative results on human motion primitives, as far as we know. Here performance is evaluated for 4 actions of the arms (gestures), namely *Point right*, *Raise arm*, *Clap* and *Wave*. The authors perform two tests, one without noise in the start and end frames of the primitives and one where the primitives are affected by noise. In the noise-free case their overall accuracy is 94.4% while in the presence of noise the accuracy is 86.9%. Our results are not immediately comparable with the ones of (Holte et al., 2010) since we use public datasets (see above Section 7.7.1, while they have built their own dataset, which is not publicly available. Furthermore, we have obtained by our classification process 16 primitives for each arm which are in accordance with biomechanics primitives. This notwithstanding, we mapped their 22 primitives, denoted by the letters $A, \ldots, V$ to our defined primitives of the groups of *Left arm* and *Right arm* (see Table 7.5). To maintain the use of public datasets we have extracted videos from our reference datasets (see above Section 7.7.1) to obtain the 4 above mentioned gestures from 10 different subjects. Hence, we have computed the motion primitives recognition accuracy on these video sets, to compare with (Holte et al., 2010). The results are shown in Table 7.5.

Table 7.5. Comparison with the 22 motion primitives of (Holte et al., 2010)

		Shoulder abd.	Shoulder add.	Elbow ext.	Elbow flex.	Shoulder Int. Rot. and elbow flex.	Shoulder Ext. Rot. and elbow ext.	Elbow Upper med. flex.	Elbow Upper med ext.
A,B,C	Point	92.3	**96.8**						
D,E,F	right	(89.6)	(93.5)						
		82.5							
G,H,I	Raise			84.5	77.5				
J,K,L	arm			(81.4)	(73.6)				
				87.5					
M,N,O	Clap					**91.7**	89.2		
P,Q,R						(87.6)	(85.9)		
						90.0			
S,T	Wave							85.4	**87.7**
U,V								(81.3)	(82.9)
								87.5	

In Table 7.5 the capital letters in the first column indicate the primitives in the language of (Holte et al., 2010). In the second column are listed the actions formed by the primitives indicated in the first column. In the first row are indicated the primitive taken from our biomechanics language, which we mapped on the (Holte et al., 2010) primitives. Results are on the diagonal, in

gray the results of (Holte et al., 2010). We have indicated in parentheses the values illustrated in the confusion matrices. While the values in the confusion matrices were mean precision averages over all experiments for all actions in all the considered datasets, here the results are with respect to an amount of videos comparable to the experiments of (Holte et al., 2010), hence they are significantly better for the indicated primitives. Despite the results are not quite comparable since we have measured our results on public databases, and in 3D, we can observe that our approach outperforms in all but one case the results in (Holte et al., 2010).

7.7.7 Discussion

The results show that our framework discovers and recognizes motion primitives with high accuracy with respect to the manually defined baseline while providing competitive results with respect to (Holte et al., 2010), the only work, to the best of our knowledge, providing quantitative results on similarly defined motion primitives.

Additionally, given the importance of studying human motion in a wide spectrum of research fields, ranging from robotics to bioscience, we believe that the human motion primitives dataset will be particularly useful in exploring new ideas and for enriching knowledge in these areas.

7.8 An application of the motion primitives model to surveillance videos

In this section we show how to set up an experiment by using motion primitives. In particular, the application we have chosen is the detection in surveillance videos of dangerous human behaviors. To set up the experiment we consider videos of anomalous and dangerous behaviors, and prove that idiosyncratic primitives, among those identified in Figure 7.12, appear to characterize these behaviors. The application is quite interesting because it highlights how the combination of primitives allows to detect specific human behaviors. On the one side the motion primitives are used for detection and on the other side they can be used also for characterizing classes of actions or classes of activities.

7.8.1 Related works and datasets on abnormal behaviors

There is a significant amount of literature on *abnormality* detection in surveillance videos. Only few of them, though, are concerned with dangerous behaviors. These methods can be further divided into those detecting dangerous crowd behaviors, in which the individual motion is superseded by large flows

as in (Mohammadi et al., 2016; Mohammadi et al., 2015; Mousavi et al., 2015; Hassner et al., 2012), and those detecting closer dangerous human behaviors.

Among the latter there are methods focusing on fights (Gracia et al., 2015), methods specialized on violence (Zhou et al., 2018; Gao et al., 2016; Deniz et al., 2014; Xu et al., 2014), on aggressive behaviors (Kooij et al., 2016), and on crime (Sultani et al., 2018). A review on methods for detecting abnormal behaviors, taking into account some of the above mentioned ones, and also discussing available datasets, is provided in (Mabrouk and Zagrouba, 2018).

In the last years, also due to the above studies, a number of datasets have been created from real surveillance videos, or from movies repositories. The most used ones are *UCSD Anomaly* (Mahadevan et al., 2010), *Avenue Dataset* (Lu et al., 2013), the *Behave* (Blunsden and Fisher, 2010) dataset, the *Violent Flows* dataset (Hassner et al., 2012), the *Hockey Fight Dataset* (Nievas et al., 2011), the *Movies Fight Dataset* from (Nievas et al., 2011) too and, finally, the recent *UCF-crime* introduced by (Sultani et al., 2018). To these datasets some authors, studying abnormal behaviors in surveillance videos, have added specific activities from *UCF101* (Soomro et al., 2012).

To detect dangerous behaviors we considered four of the above datasets most suitable for the task of analyzing human behaviors with small groups of subjects. The first dataset is the *Hockey Fight Dataset* provided by (Nievas et al., 2011), which is formed by 1000 clips of actions from hockey games of the National Hockey League (NHL). A second dataset, also introduced by (Nievas et al., 2011) is the *Movies Fight dataset*, which is composed of 200 video clips obtained from action movies, 100 of which show a fight. Videos in both these datasets are untrimmed but divided in those where there are fights and those where there are no fights. The third dataset is the *UCF-Crime dataset* introduced by (Sultani et al., 2018). This dataset is formed by 1900 untrimmed surveillance videos of 13 realworld anomalies, including *abuse, arrest, arson, assault, road accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, and vandalism*, and normal videos. These videos have varying length from 30 sec. up to several minutes. In a number of these videos, like explosion and road accident, no human behavior is observable. Among the others there are a number of videos not including human behaviors. Therefore we have chosen a subset of all the UCF-crime dataset for both training and testing. In particular, we have chosen abuse, arrest, assault, burglary, fighting, robbery, shooting, stealing, and vandalism. Finally we have taken videos from *UCF101* dataset, which includes 101 human activities.

Given the above selected datasets we aim at showing that once the primitives are computed an off-the-shelf classifier can be used to detect specific behaviors, in this case the dangerous ones.

The method we propose requires to compute the primitives on a selected training set, separating the untrimmed videos with dangerous behaviors from the normal ones, as described below, and then training a non-linear kernel SVM on the two datasets, as illustrated in Section 7.8.3. The trained classifier is

then tested on the test sets and results are reported in Section 7.8.4, comparing with state of the art approaches.

The main idea we want to convey here is that once primitives are computed all the relevant features for distinguishing a behavior are embedded in the primitive category of the specific group (see Section 7.8.4) and therefore the classifier has to deal just with them and not with other features such as poses, images, time and tracking, in so alleviating the classifier burden and allowing to deal with state of the art classifiers. Furthermore, the primitive parameters, used to estimate the primitive classes, are no more needed for the further classification of behaviors. This is the main advantage of human motion primitives modeling, namely their effectiveness in characterizing specific behaviors.

7.8.2 Primitives computation

For primitives computation we collected all the videos from hokey and fight-movie datasets, we collected from the UCF-crime dataset the videos from *abuse, arrest, assault, burglary, fighting, robbery, shooting, stealing*, and *vandalism*. Finally, from UCF101 we collected 276 videos from the datasets *Punch* and *SumoWrestling* and further 276 videos from other sports, randomly chosen as in (Gracia et al., 2015). The total number of videos collected is 3050 for primitive computation, as illustrated in the following table:

Table 7.6. Datasets for primitive computation in dangerous behaviors detection

	Hockey		Fight-Movies		UCF-crime		UCF101	
	Danger.	Normal	Danger.	Normal	Danger.	Normal	Danger.	Normal
Video sets	500	500	100	100	650	650	276	276
Training	70%	70%	70%	70%	70%	70%	100%	70%
Test	30%	30%	30%	30%	30%	30%	0%	30%

To compute the primitives for each subject from a small group of people appearing in a frame of a video, we have fitted 3D poses basing on the SMPL model (Loper et al., 2015) of *human mesh recovery* (HMR) (Kanazawa et al., 2018a). HMR recovers together with joints and pose also a full 3D mesh from a single image (see Figures 7.15 and 7.16), and it is accurate enough to estimate multiple subject poses in a single frame.

Having more than a subject requires to track each subject pose across frames, in order to compute the motion primitives for each of them. To this end we used the joints given by SMPL model in world frame, for the following body joints (see the preliminary Section 7.4): left and right *hip*, left and right *clavicle* (called shoulder in HMR), and the *head*. These joints are well suited for tracking since they have slower motion with respect to other body parts. Tracking amounts to find the rotations and translations amid all the bodies appearing in two consecutive frames, and identifying the rotation

and translation pertaining to each subject across the two frames. Consider two consecutive frames indexed by t and t+1, and let $\mathcal{J}^{(t)} = \{j_1^{(t)}, \ldots, j_5^{(t)}\}$ and $\mathcal{J}'^{(t+1)} = \{j'^{(t+1)}_1, \ldots, j'^{(t+1)}_5\}$ be the joints in world frame of the above mentioned body components, where joint subscripts indicate in the order left and right hip, left and right clavicle and head. We first find the translation $\mathbf{d}$ and rotation R between any two set of joints appearing in the frames t and t+1 (see also Section 7.4):

$$(R, \mathbf{d}) = \underset{R \in SO(3), \mathbf{d} \in \mathbb{R}^3}{\arg\min} \sum_{i=1}^{5} w_i \| (R\ j_i^{(t)} + \mathbf{d}) - j'^{(t+1)}_i \|^2 \tag{7.19}$$

With $w_i > 0$ weights for each pair of joints in (t) and $(t+1)$. Let $\hat{\mathcal{J}} = (\sum_{i=1}^{5} w_i j_i) / \sum_{i=1}^{5} w_i$ be the weighted centroids of the set of joints $\mathcal{J}$. The minimization in (7.19) is solved by computing the singular value decomposition $U\Sigma V^\top$ of the covariance matrix $\bar{\mathcal{J}}^{(t)} W (\bar{\mathcal{J}}'^{(t+1)})^\top$ of the normalized joints $\bar{\mathcal{J}}^{(t)}, \bar{\mathcal{J}}'^{(t+1)}$, obtained by subtracting the weighted centroid to each joints set. Here W is the diagonal matrix of the weights w_i. Let H be the diagonal matrix $diag(\mathbf{1}, det(VU^\top))$, then the rotations and translations between sets of joints are found as:

$$R = VHU^\top \quad \text{and } \mathbf{d} = \hat{\mathcal{J}}'^{(t+1)} - R\hat{\mathcal{J}}^{(t)} \tag{7.20}$$

Finally, once we have obtained the rotation matrices and the translation vectors between the sets of considered joints of all the fitted skeletons, from frame t to frame $t+1$, we can track each individual skeleton S_k. A skeleton $S_k^{(t+1)}$ belongs to the same subject fitted by skeleton $S_k^{(t)}$, at frame t, if the rotation R_k and translation $\mathbf{d}_k$, obtained according to eq. (7.20) between the chosen joints $\mathcal{J}^{(t)}$ of $S_k^{(t)}$ and $\mathcal{J}'^{(t+1)}$ of $S_k^{(t+1)}$, satisfy

$$(R_k, \mathbf{d}_k) = \underset{R_k \in SO(3), \mathbf{d}_k \in \mathbb{R}^3, k=1:s}{\arg\min} \| \mathcal{J}'^{(t+1)} - ((\mathcal{J}^{(t)} R_k)^\top + \mathbf{d}_k)^\top \|_F \tag{7.21}$$

With $\| \cdot \|_F$ the Frobenious norm and $s = N_S!/((N_S-2)!2!)$, with N_S the common number of fitted skeletons S in both frame t and $t+1$.

Once the skeletons are tracked we can compute the unknown primitives from the flux (see Section 7.5) as paths $\gamma^T_{G_m} : I \subset \mathbb{R} \mapsto \mathbb{R}^9$, for each group G_m, with I the time interval, specified by the frame sequence, and scale it as described in Section 7.5. We can then use the parameters Θ learned with the recognition model, detailed in Section 7.6.2, to assign a label $\mathcal{L}_w^{G_m}$ to each primitive segmented by the motion flux as precised in eq. (7.18). Namely, we find the model identified by the parameter Θ_w, which maximizes the probability of the primitive under consideration. We recall that for each group G_m, $m = 1, \ldots, 6$ there are q models with $q \in \{7, 10, 16\}$ (see the primitives representation in Figure 7.12).

Our model of motion primitives relies significantly on the accuracy of the 3D pose estimation. We have chosen the model HMR (Kanazawa et al., 2018a)

based on SMPL (Loper et al., 2015), in place of (Natola et al., 2015b; Tome et al., 2017), since it is most recent and highly accurate. Still not all the videos chosen obtain a reasonable fitting, therefore after skeleton fitting and tracking a number of videos from UCF-crime have been removed from the considered set.

7.8.3 Training a non-linear binary classifier

All the computed primitives are labeled by their name (e.g. *Elbow flex*), according to the recognition model, as specified above. A set of primitives for a given video is formed as follows. Primitive names are embedded into real numbers $r \sim Unif(0,1)$, such that for each primitive name there is a precise real number. Given frame t for each skeleton appearing in the frame we form a vector of dimension 6×1, where the 6 elements are the corresponding embedded primitive names occurring at frame t. Let $\gamma_{G_m}^{(t)}$ denote the primitive of the body group G_m, and u the mapping of the primitive name to the real number:

$$\mathbf{x}_j^{(t)} = (u(\gamma_{G_1}^{(t)}), u(\gamma_{G_2}^{(t)}), \ldots, u(\gamma_{G_6}^{(t)}))^\top \tag{7.22}$$

Where j indicates the j-th skeleton appearing in frame t. Note that t and j are actually indicated just for forming the training set, to select from all the gathered vectors $\mathbf{x}$ those that have changing primitives. Namely, for training, from the set of all vectors in each frame, we have retained only those vectors in which at least one primitive changes, for each recorded skeleton.

For training we have selected videos for both dangerous behaviors and normal behaviors, thus labeling them with 1 for dangerous and -1 for normal behaviors, as follows. We selected 70% of fighting and 70% of not fighting from both hockey and fight movies; from UCF101 we have selected all videos in *Punch* and *SumoWrestling*, getting 276 videos and further 276 videos randomly from sport activities. For UCF-crime we proceeded as follows. We have selected the videos from all the crime activities specified above with time length less than $60sec.$ and cropped the first and last $10sec.$, in order to do a weak supervised training, namely, as in (Sultani et al., 2018) we have not trimmed the video. Thus we obtained 173 videos for abnormal activities and we selected 173 videos from the normal activities. The total number of videos for training is 1634 videos. All the remaining video with computed primitives have been used for testing.

The resulting data structure is:

$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \text{ with } \mathbf{x} \in \mathbb{R}^6, y \in \{-1, 1 | \ -1 \text{ if normal, } 1 \text{ if dangerous}\} \tag{7.23}$$

The SVM (Vapnik, 2013) is a popular classification method computing, for two non-separable classes, the classifier:

$$\begin{aligned} f(\mathbf{x}) &= (\textstyle\sum_{i=1}^n y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b) \\ \hat{y} &= sgn(f(\mathbf{x})) \end{aligned} \tag{7.24}$$

where K is the kernel function $\varphi(\mathbf{x}_i)^\top \varphi(\mathbf{x}_j)$ with φ the feature map, here we considered the RBF kernel $\exp\left(-\eta \|\mathbf{x}_i - \mathbf{x}_j\|_{\ell_2}^2\right)$, with η a tunable parameter. Classification is obtained by solving the constrained optimization problem:

$$\max_{\alpha} \frac{1}{2}\boldsymbol{\alpha}^\top \Omega \boldsymbol{\alpha} - \mathbf{e}^\top \boldsymbol{\alpha} \quad \text{subject to} \quad \mathbf{y}^\top \alpha = 0,\ 0 \leq \alpha_i \leq \lambda \tag{7.25}$$

Here Ω is a square $n \times n$ positive semidefinite matrix, with $\omega_{i,j} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{e}$ is a vector of ones, the non zero α_i define the support vectors, and λ is the regularization parameter of the primal optimization problem $min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}\mathbf{w}^\top + \lambda \sum_{i=1}^{n} \xi_i$ (Scholkopf and Smola, 2001). To obtain posterior probabilities we applied the Platt scaling (Platt et al., 1999), proposing a sigmoid model to fit a posterior on the SVM output:

$$P(y = 1|f(\mathbf{x})) = \frac{1}{1 + \exp(Af(\mathbf{x}) + B)} \tag{7.26}$$

Here the parameters A and B are fitted by solving the maximum likelihood problem:

$$min_{z=(A,B)} F(z) = -\sum_{i=1}^{n} \left(t_i \log(p_i) + (1 - t_i) \log(1 - p_i)\right) \tag{7.27}$$

Using as prior the number of positive N_+ and negative N_- examples in the training data, with $p_i = P(y = 1|f(\mathbf{x}_i))$, $t_i = (N_+ + 1)/(N_+ + 2)$ if $y_i = 1$ and $1/(N_- + 2)$ if $y_i = -1$. See also (Lin et al., 2007) for an improved algorithm with respect to (Platt et al., 1999).

To obtain the probability that at a given frame t a dangerous event occurs we compute the average response to the primitives of each subject which has been detected. More precisely, let s be the number of subjects in frame t for which the primitives are computed, then the observation $\mathbf{x}^{(t)} = (\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_s^{(t)})$. Given $\mathbf{x}^{(t)}$, and assuming that the SVM scores for each $\mathbf{x}_i^{(t)}$ are independent, we can define the probability that a dangerous event Y is occurring at t, in a surveillance video, as the expectation:

$$P(Y|\mathbf{x}^{(t)}) = \sum_{i=1}^{s} p(\hat{y}_i^{(t)}|\mathbf{x}_i^{(t)}) P(y_i = 1|f(\mathbf{x}_i^{(t)})) \tag{7.28}$$

Here $p(\hat{y}^{(t)}|\mathbf{x}^{(t)})$ is computed by remapping the scores to $[0, 1]$ such that $\sum_{i=1}^{s} p(\hat{y}_i^{(t)}|\mathbf{x}_i^{(t)}) = 1$. Testing has been done on the videos on which the primitives have been precomputed, and the results are shown together with comparisons with the state of the art in Section 7.8.4. Note that the method is not yet suitable for online detection of dangerous behaviors, still it can be advanced to online detection, by lifting the computation of the flux with motion anticipation.

Table 7.7. AUC comparison with state-of-the-art methods on the UCF-Crime dataset.

Method	Binary classifier	Hasan et al. (Hasan et al., 2016)	Lu et al. (Lu et al., 2013)	(Sultani et al., 2018)	(Sultani et al., 2018) w. constraints	Ours
AUC	50.0	50.6	65.51	74.44	75.41	**76.15**

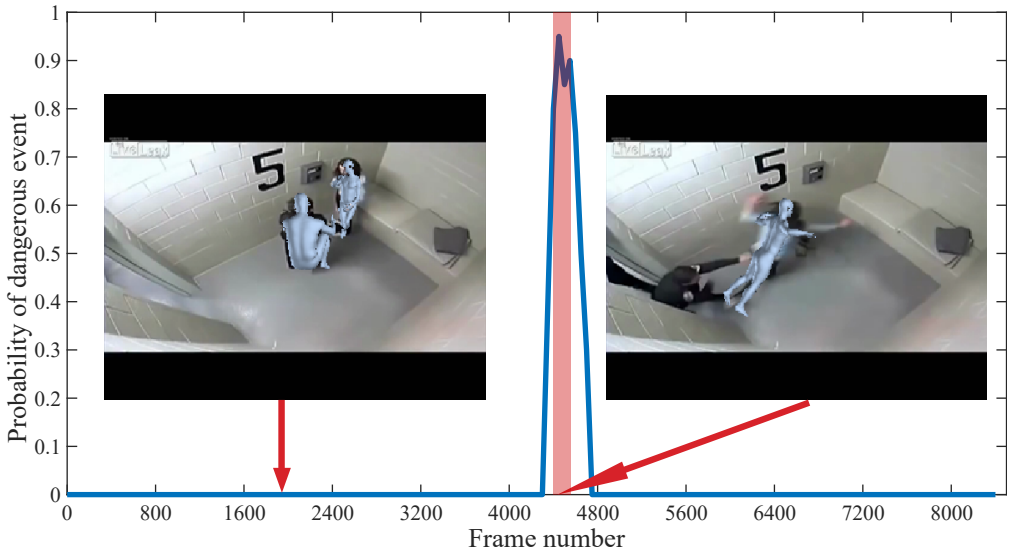
7.8.4 Results and comparisons with the state of the art

We discuss now the results achieved by our method for abnormal behavior detection based on human motion primitives. Figure 7.15 shows some qualitative results of dangerous behaviors detection in four videos. Three videos correspond to crime activities, namely *Abuse*, *Fighting* and *Shooting*, while the last displays a normal activity. The curve plotted in the graphs provides for each frame the probability that a dangerous event is occurring, according to eq. (7.28). The highlighted region corresponds to the interval where a crime activity occurs. From this graphs it is evident that the crime activity detection follows closely the ground truth. For each example we also show two representative frames overlaid with the human meshes identified by HMR. Similarly, Figure 7.17 shows some representative examples of fitted human meshes for videos taken from Hockey and Movie Fights datasets.

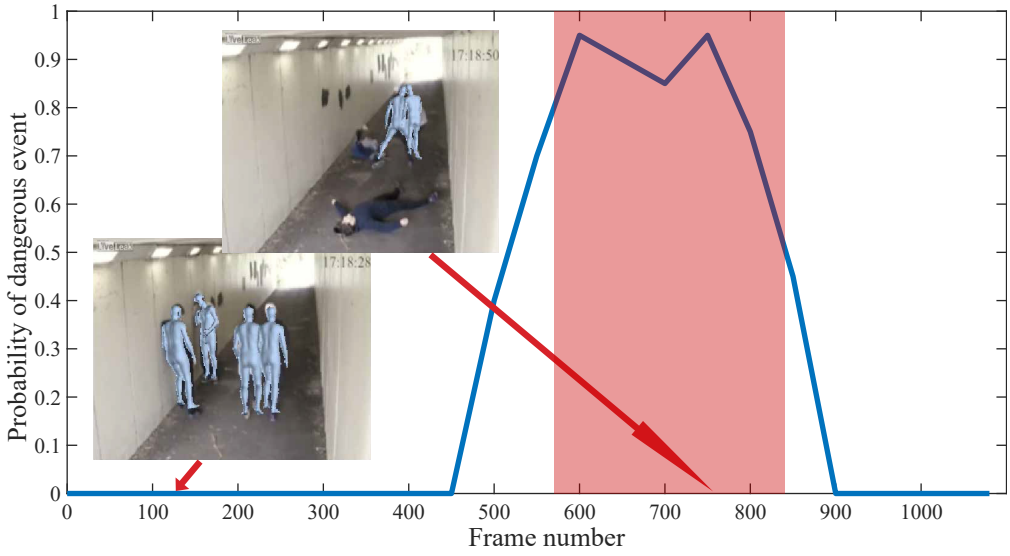
Fig. 7.19 presents the ROC curves of the proposed method for the four datasets considered, namely UCF-Crime, UCF101, Hockey Fights and Movie Fights. The corresponding values of the area under curve (AUC) are 76.15%, 91.92%, 98.44% and 98.77%, respectively. Table 7.8 presents the mean accuracy, its standard deviation and the area under the receiver-operating-characteristic (ROC) curve of our method in comparison with other state-of-the-art methods. The results of the other methods are taken from (Gracia et al., 2015). We observe that our method achieves better performance on the Hockey Fights and Movies Fights datasets while it has very similar performance with the best performing method on the UCF101 dataset.

Additionally, in Figure 7.18 we present the frequency graphs of primitive occurrences for groups G2 and G3, for the crime activities *Abuse*, *Fighting*, *Robbery*, and *Shooting*. The graphs show that each type of activity manifests itself by a different combination of idiosyncratic motions of the limbs. This fact can be used to achieve finer grained categorization of the crime activities, however, we do not examine further this possibility in this work.

Figure 7.19 presents the ROC curves of the proposed method for the four datasets considered, namely UCF-Crime, UCF101, Hockey Fights and Movie Fights. The corresponding values of the area under curve (AUC) are 76.15%, 91.92%, 98.44% and 98.77%, respectively. Table 7.8 presents the mean accuracy, its standard deviation and the area under the receiver-operating-characteristic (ROC) curve of our method in comparison with other state-of-the-art methods. The results of the other methods are taken from (Gracia et al., 2015). We observe that our method achieves better performance on the Hockey Fights and Movies Fights datasets while it has very similar performance with the best

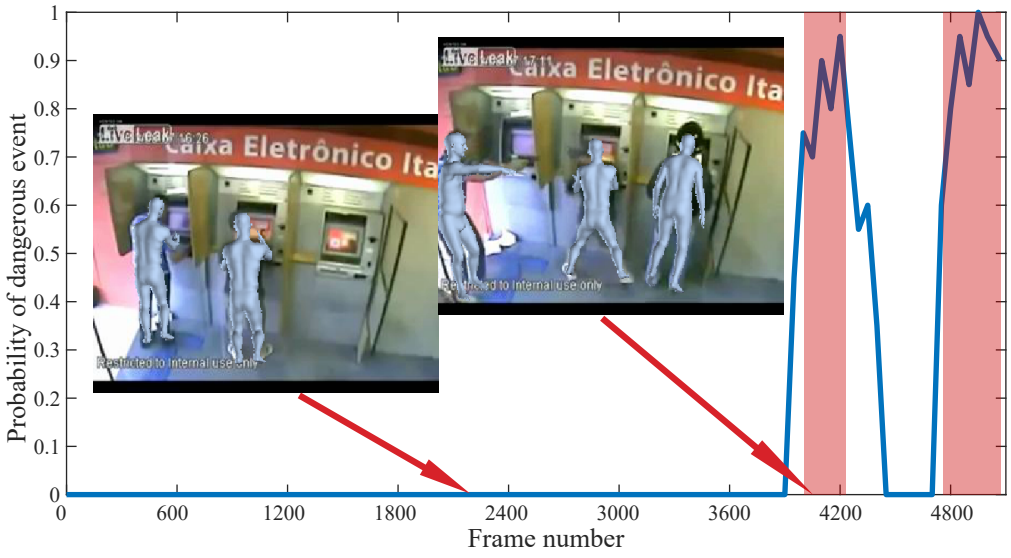


(a)

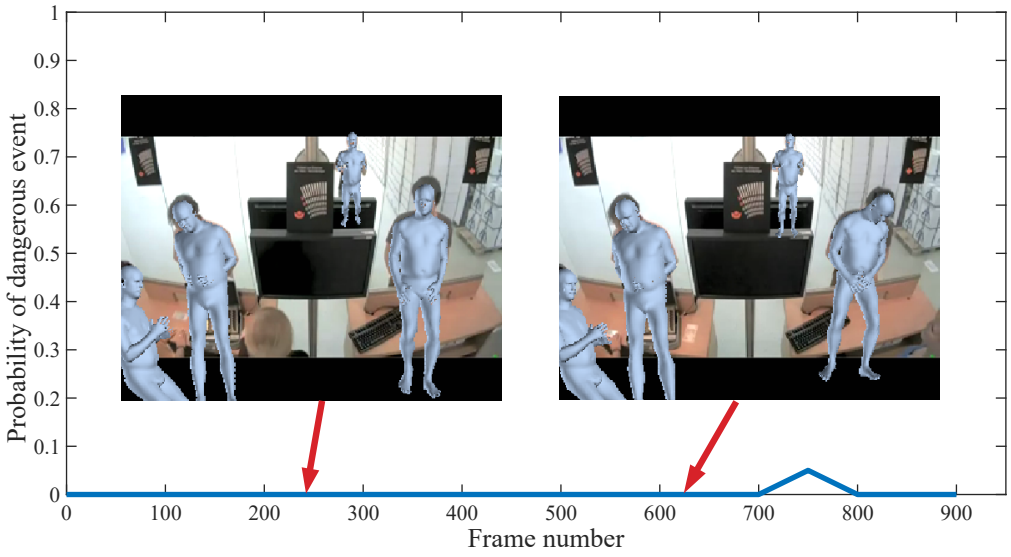


(b)

Figure 7.15. Results of the proposed method on videos from UCF-Crime dataset. From top: *Abuse*, *Fighting*. Colored window shows ground truth anomalous region.

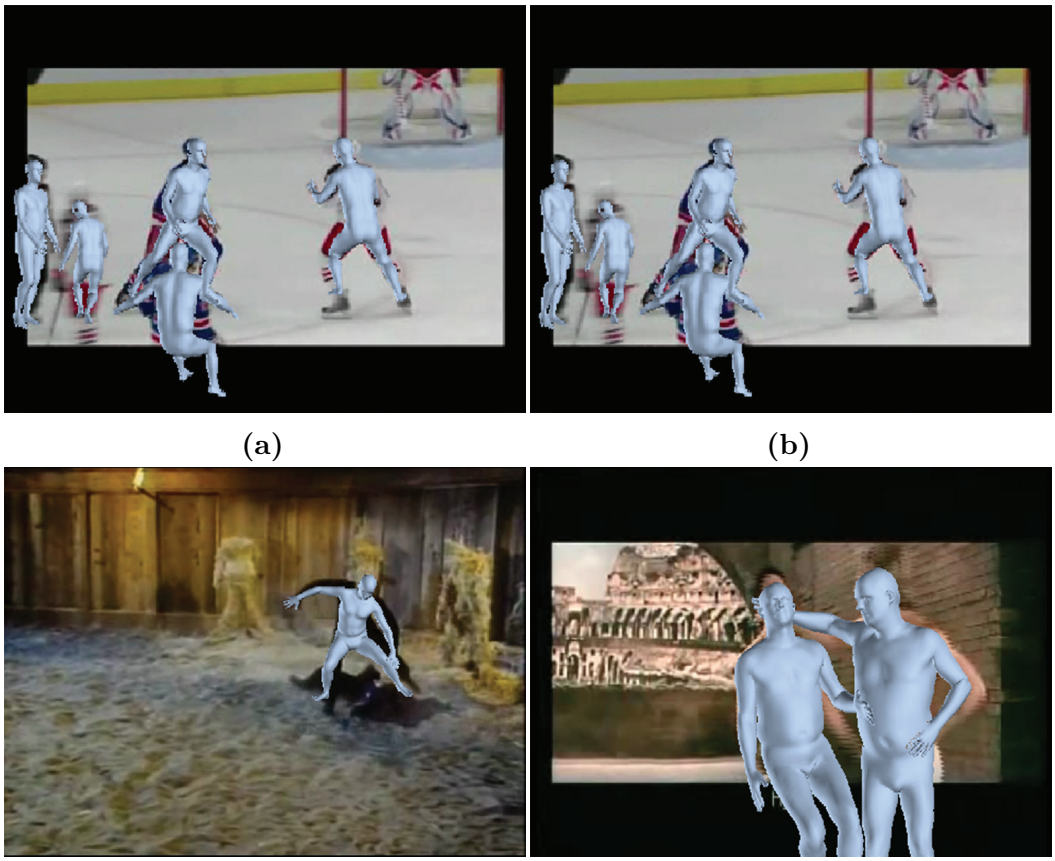


(a)



(b)

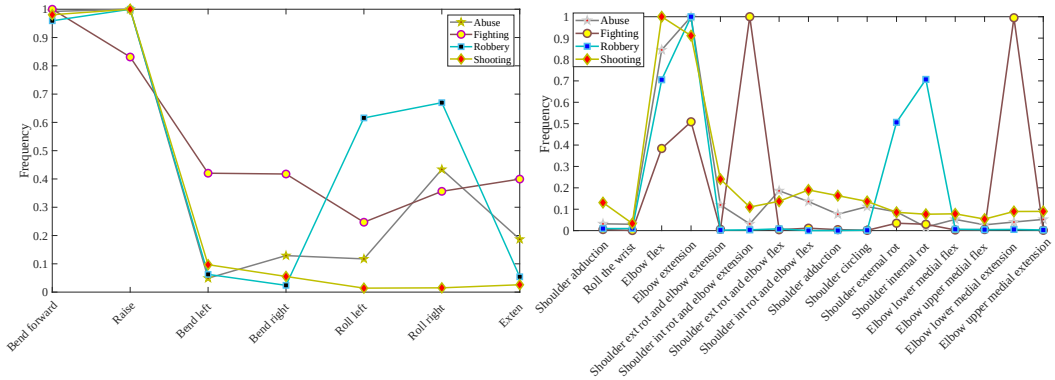
Figure 7.16. Results of the proposed method on videos from UCF-Crime dataset. From top: *Shooting*, *Normal*. Colored window shows ground truth anomalous region.



(a) **(b)**

(c) **(d)**

Figure 7.17. Instances of videos with human meshes fitted using HMR from Hockey and Movies datasets (Nievas et al., 2011).



(a) Primitives of *torso*, group G_2. **(b)** Primitives of *right arm*, group G_3.

Figure 7.18. Frequency graphs of the occurrences of primitives for groups G_2 (torso) and G_3 (right arm) in the videos of *Abuse*, *Fighting*, *Robbery*, and *Shooting*of the dataset UCF-crime.

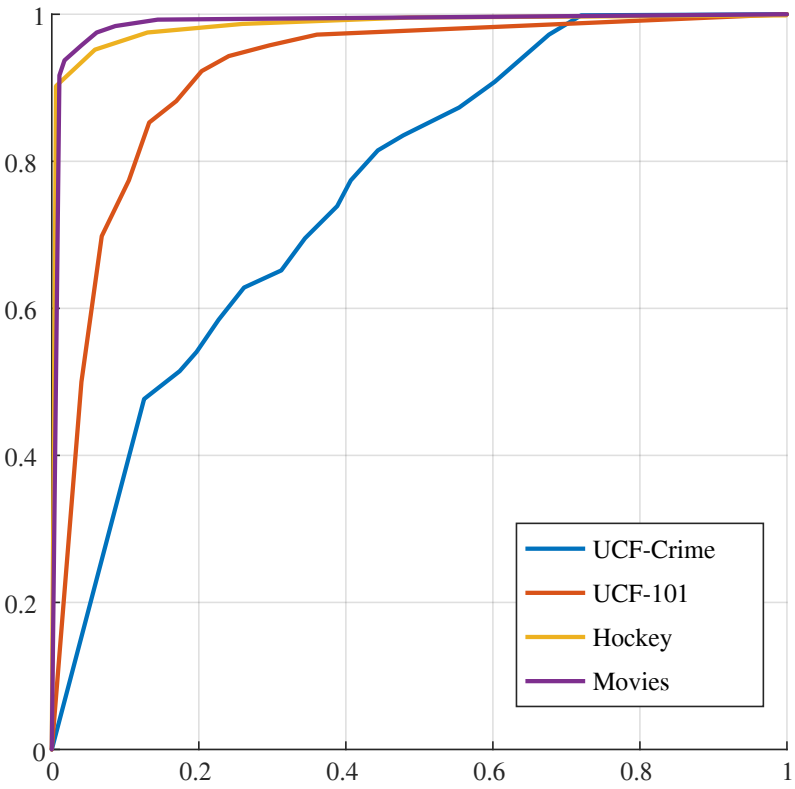


Figure 7.19. ROC curves of the proposed method for UFC-Crime, UFC101, Hockey and Movies datasets.

performing method on the UCF101 dataset.

Finally, Table 7.7 gives a comparison of the results achieved by our method on the UCF-Crime dataset in comparison with results from other state-of-the-art methods as reported in (Sultani et al., 2018). In this case we have to highlight that our results are not directly comparable with the ones reported in (Sultani et al., 2018) as we restrict our analysis on videos where human subjects are visible. Nevertheless, the results indicate that also on this database the proposed method is able to achieve state-of-the-art performance on crime activity detection.

7.9 Conclusions

We presented a framework for automatically discovering and recognizing human motion primitives from video sequences based on the motion of groups of joints of a subject. To this end the motion flux is introduced which captures the variation of the velocity of the joints within a specific interval. Motion primitives are discovered by identifying intervals between rest instances that maximize the motion flux. The unlabeled discovered primitives have been separated into different categories using a non-parametric Bayesian mixture model.

Table 7.8. Comparison with state-of-the-art methods on the datasets Movies, UCF101 and Hockey.

Method	Classifier	Datasets		
		Movies	**Hockey**	**UCF101**
BoW (STIP)	SVM	82.3±0.9/0.88	88.5±0.2/0.95	72.5±1.5/0.74
	AdaBoost	75.3±0.83/0.83	87.1±0.2/0.93	63.1±1.9/0.68
	RF	97.7±0.5/0.99	96.5±0.2/0.99	87.3±0.8/0.94
BoW (MoSIFT)	SVM	63.4±1.6/0.72	83.9±0.6/0.93	81.3± 1/0.86
	AdaBoost	65.3±2.1/0.72	86.9±1.6/0.96	52.8±3.6/0.62
	RF	75.1±1.6/0.81	96.7±0.7/0.99	86.3±0.8/0.93
ViF	SVM	96.7±0.3/0.98	82.3±0.2/0.91	77.7±2.16/0.87
	AdaBoost	92.8±0.4/0.97	82.2±0.4/0.91	78.4±1.7/0.86
	RF	88.9±1.2/0.97	82.4±0.6/0.9	77±1.2/0.85
LMP	SVM	84.4±0.8/0.92	75.9±0.3/0.84	65.9±1.5/0.74
	AdaBoost	81.5±2.1/0.86	76.5±0.9/0.82	67.1±1/0.71
	RF	92±1/0.96	77.7±0.6/0.85	71.4±1.6/0.78
(Deniz et al., 2014)	SVM	85.4±9.3/0.74	90.1±0/0.95	**93.4±6.1/0.94**
	AdaBoost	98.9±0.22/0.99	90.1±0/0.90	92.8±6.2/0.94
	RF	90.4±3.1/0.99	61.5±6.8/0.96	64.8±15.9/0.93
(Gracia et al., 2015) v1	SVM	87.9±1/0.97	70.8±0.4/0.75	72.1±0.9/0.78
	AdaBoost	81.8±0.5/0.82	70.7±0.2/0.7	71.7±0.9/0.72
	RF	97.7±0.4/0.98	79.3±0.5/0.88	74.8±1.5/0.83
(Gracia et al., 2015) v2	SVM	87.2±0.7/0.97	72.5±0.5/0.76	71.2±0.7/0.78
	AdaBoost	81.7±0.2/0.82	71.7±0.3/0.72	71±0.8/0.72
	RF	97.8±0.4/0.97	82.4±0.6/0.9	79.5±0.9/0.85
Ours	SVM	**99.1±0.3/0.99**	**97.2±0.8/0.98**	93.3±2.1/0.92

We experimentally show that each primitive category naturally corresponds to movements described using biomechanical terms. Models of each primitive category are built which are then used for primitive recognition in new sequences. The results show that the proposed method is able to robustly discover and recognize motion primitives from videos, by using state-of-the-art methods for estimating the 3D pose of the subject of interest. Additionally, the results suggest that the motion primitives categories are highly discriminative for characterizing the activity been performed by the subject.

Finally, a dataset of motion primitives is made publicly available to further encourage result reproducibility and benchmarking of methods dealing with the discovery and recognition of human motion primitives.

Chapter 8

Conclusions: Implications and Future Directions

In this concluding chapter we summarize the contributions of this thesis, and discuss the possible impacts and the important directions of future work.

8.1 Summary of thesis contributions

Problems related to three main computer vision areas were addressed in this thesis: 3D object modeling from single or multiple images, 3D human pose estimation from single images and human motion analysis from RGB videos.

We proposed a method for computing 3D models of articulated objects from few multiple images, by decomposing them into components. Realistic models of the object components were built by merging together 3D models obtained from different aspects. The entire object was obtained by reassembling the components using two or more images of the object in a reference pose. Furthermore, software code for this paper was made available at `https://github.com/alcor-lab/articulated-object-modeling`.

We proposed an approach for BRDF aware modeling of 3D objects from a single image. We were able to fully model non-Lamberitan surfaces with either concave or sharp parts and we have proved that the normal field of the surfaces to be modeled can be learned from renderings of different objects surfaces.

We presented a method for 3D human pose estimation from a single image based on a hierarchical Bayesian non-parametric model. The proposed model captures variations of the motion and the appearance of different body parts, identified by groups of human skeleton joints. The decomposition in groups avoids redundant configurations, obtaining a more concise dictionary of poses and visual appearances.

We presented a framework for automatically discovering and recognizing human motion primitives from RGB video sequences based on the motion of groups of 3D skeleton joints of a subject. The motion flux is introduced which

captures the variation of the velocity of the joints within a specific interval. The unlabeled discovered primitives have been separated into different categories using a non-parametric Bayesian mixture model. Each primitive category naturally corresponds to movements described using biomechanical terms. Models of each primitive category are built which are then used for primitive recognition in new sequences. The results show that the proposed method is able to robustly discover and recognize motion primitives from videos, by using state-of-the-art methods for estimating the 3D pose of the subject of interest. Additionally, the results suggest that the motion primitives categories are highly discriminative for characterizing the activity been performed by the subject. Furthermore, a dataset of motion primitives is made publicly available at `https://github.com/alcor-lab/MotionPrimitives`.

8.2 Direction for future work

The natural future direction will be to build a framework for activity recognition from RGB videos based on context. The available state of the art research still misses a complete framework for human activity recognition based on context, taking into account both the scene where activities are taking place, objects analysis, 3D human motion analysis and interdependence between activity classes. This thesis describes computer vision frmeworks which will enable the robust recognition of human activities explicitly considering the scene context.

The main contribution will be to consider an human activity in context, both taking into account the dynamic relations a person is carrying on with objects in the scene and other related objects, which can be more or less relevant. This information will boosts the recognition accuracy of complex activities, which would be otherwise hard in videos such as those taken from the web, for example those collected in both ActivityNet (Fabian Caba Heilbron and Niebles, 2015) and (Kuehne et al., 2011).

To successfully recognize an activity in a video some minimal conditions need to be satisfied, these are: revealing a cause effect relation connecting a subject pose and some object in the scene, time persistence of this relation and the recognition of other relevant objects that allude to the activity in course. To comply with these conditions, relevant features can be computed related to objects, to subjects pose and features enabled by tracking the poses and the interactions between subjects-objects and amid relevant objects. The research conducted so far in 3D objects modeling, 3D human pose estimation and 3D human motion analysis will enable the discovery of such relevant features.

In order to detect objects in the scene, very popular neural networks can be used, such as the Faster R-CNN network (Ren et al., 2015), delivering the object class together with its bounding box. Furthermore, features enabled from the 3D shape of objects can be used (Ntouskos et al., 2015b; Natola et al., 2016). To estimate the 2D pose of the subjects in the video, Realtime

Multi-person network (RTMPOSE) (Cao et al., 2016) or the popular OpenPose network (Cao et al., 2018) can be used, providing an estimation of the 2D poses of multiple subjects appearing in the scene. To verify persistence of objects prediction in a time lapse the T-CNN tracking method (Kang et al., 2016) can be used, building on a temporal convolutional network that operates on a spatio-temporal object proposal. To recover 3D human pose from 2D pose, methods such as (Sanzari et al., 2016; Pavllo et al., 2019; Kanazawa et al., 2018b) can be used. Finally, to analyze 3D human motion we can use (Sanzari et al., 2019).

The proposed method will allow to not only recognize the activity but to identify also the subjects and the objects involved in the specific activity. Furthermore, incorporating information regarding the nature of the scene will further assist the recognition process. It would be crucial to use the information that an activity takes place outdoors, indoors in a rural or in an urban scene in order to robustly discriminate between a wide range of activities.

Bibliography

CMU Mocap Database. http://mocap.cs.cmu.edu/.

https://renderpeople.com/3d-people/.

https://www.turbosquid.com/.

Abernethy, B. (2013). *Biophysical foundations of human movement*. Human Kinetics.

Afsari, B., Tron, R., and Vidal, R. (2013). On the convergence of gradient descent for finding the riemannian center of mass. *SIAM Journal on Control and Optimization*, 51(3):2230–2260.

Agarwal, A. and Triggs, B. (2006). Recovering 3d human pose from monocular images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(1):44–58.

Akhter, I. and Black, M. J. (2015a). Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1446–1455.

Akhter, I. and Black, M. J. (2015b). Pose-conditioned joint angle limits for 3d human pose reconstruction. In *CVPR*, pages 1446–1455.

Alt, H. and Guibas, L. J. (2000). Discrete geometric shapes: Matching, interpolation, and approximation. In *Handbook of computational geometry*, pages 121–153. Elsevier.

Amor, H. B., Neumann, G., Kamthe, S., Kroemer, O., and Peters, J. (2014). Interaction primitives for human-robot cooperation tasks. In *ICRA*, pages 2831–2837.

Andrew, G. and Gao, J. (2007). Scalable training of l1-regularized log-linear models. In *ICML*, pages 33–40.

Andriluka, M., Roth, S., and Schiele, B. (2010). Monocular 3d pose estimation and tracking by detection. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 623–630. IEEE.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Asfour, T., Gyarfas, F., Azad, P., and Dillmann, R. (2006). Imitation learning of dual-arm manipulation tasks in humanoid robots. In *International Conference on Humanoid Robots*, pages 40–47.

Aspert, N., Santa Cruz, D., and Ebrahimi, T. (2002). Mesh: measuring errors between surfaces using the hausdorff distance. In *ICME*, pages 705–708.

Azad, P., Asfour, T., and Dillmann, R. (2007). Toward an unified representation for imitation of human motion on humanoids. In *Robotics and Automation*, pages 2558–2563.

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., and Baskurt, A. (2010). Action classification in soccer videos with long short-term memory recurrent neural networks. In *International Conference on Artificial Neural Networks*, pages 154–159. Springer.

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., and Baskurt, A. (2011). Sequential deep learning for human action recognition. In *International workshop on human behavior understanding*, pages 29–39. Springer.

Barron, J. and Malik, J. (2015). Shape, illumination, and reflectance from shading. *TPAMI*.

Barrow, H. and Tenenbaum, J. (1978). Recovering intrinsic scene characteristics from images. *Computer Vision Syst.*

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *ECCV*, pages 404–417. Springer.

Behera, A., Hogg, D. C., and Cohn, A. G. (2012). Egocentric activity monitoring and recovery. In *Asian Conference on Computer Vision*, pages 519–532. Springer.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(4):509–522.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *TPAMI*, 35(8):1798–1828.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115.

Biess, A., Liebermann, D. G., and Flash, T. (2007). A computational model for redundant human three-dimensional pointing movements: Integration of independent spatial and temporal motor plans simplifies movement dynamics. *J. Neuroscience*, 27(48):13045–13064.

Billard, A. G., Calinon, S., and Guenter, F. (2006). Discriminative and adaptive imitation in uni-manual and bi-manual tasks. *Robotics and Autonomous Systems*.

Binford, T. O. (1971). Visual perception by computer. In *Conference on Systems and Control*, volume 261, page 262. IEEE.

Bizzi, E. and Mussa-Ivaldi, F. A. (1995). Toward a neurobiology of coordinate transformations. In *The Cognitive Neurosciences*, pages 495–506.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for dirichlet process mixtures. *Bayes. Anal.*, 1(1):121–143.

Blunsden, S. and Fisher, R. (2010). The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4(1-12):4.

Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proc. of the Int'l Conf. on Image and Video Retrieval*, pages 401–408. ACM.

Botsch, M. and Sorkine, O. (2008). On linear variational surface deformation methods. *Visualization and Computer Graphics, IEEE Transactions on*, 14(1):213–230.

Bouchard, D. and Badler, N. (2007). *Semantic Segmentation of Motion Capture Using Laban Movement Analysis*, pages 37–44. Springer.

Bredies, K., Kunisch, K., and Pock, T. (2010). Total generalized variation. *SIAM JIS*, 3(3):492–526.

Budd, C., Huang, P., Klaudiny, M., and Hilton, A. (2013). Global non-rigid alignment of surface sequences. *IJCV*, 102(1-3):256–270.

Burger, M. and Osher, S. (2013). A guide to the tv zoo. In *Level Set and PDE Based Reconstruction Methods in Imaging*, pages 1–70. Springer.

Caddigan, E., Choo, H., Fei-Fei, L., and Beck, D. M. (2017). Categorization influences detection: A perceptual advantage for representative exemplars of natural scene categoriescaddigan, choo, fei-fei, & beck. *Journal of Vision*, 17(1):21–21.

Calakli, F. and Taubin, G. (2011). Ssd: Smooth signed distance surface reconstruction. *Pacific Graphics*.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

Cao, Z., Simon, T., Wei, S., and Sheikh, Y. (2016). Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*.

Carreira, J., Kar, A., Tulsiani, S., and Malik, J. (2014). Virtual view networks for object reconstruction. *arXiv preprint arXiv:1411.6091*.

Cashman, T. J. and Fitzgibbon, A. W. (2013). What shape are dolphins? building 3d morphable models from 2d images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):232–244.

Celniker, G. and Gossard, D. (1991). Deformable curve and surface finite-elements for free-form shape design. In *ACM SIGGRAPH computer graphics*, volume 25, pages 257–266. ACM.

Chambolle, A. and Pock, T. (2010). A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *JMIV*, 40(1):120–145.

Chandraker, M. K., Kahl, C. F., and Kriegman, D. J. (2005). Reflections on the generalized bas-relief ambiguity. In *CVPR*, volume 1, pages 788–795.

Chang, J. and Fisher III, J. W. (2013). Parallel sampling of dp mixture models using sub-cluster splits. In *Advances in Neural Information Processing Systems*, pages 620–628.

Chen, T., Zhu, Z., Shamir, A., Hu, S.-M., and Cohen-Or, D. (2013). 3-sweep: Extracting editable objects from a single photo. *ACM TOG*, 32(6):195.

Cheron, G., Laptev, I., and Schmid, C. (2015). P-cnn: Pose-based cnn features for action recognition. In *ICCV*.

Cipolla, R. (1998). The Visual Motion of Curves and Surfaces. *Phil. Trans. Royal Soc. London A*, 356:1103–1121.

Cipolla, R. and Giblin, P. (2000). *Visual Motion of Curves and Surfaces*. Cambridge.

Csiszár, I. (1996). Maxent, mathematics, and information theory. In *Max. entropy and Bayesian methods*, pages 35–50. Springer Science & Business Media.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE.

de los Reyes-Guzmán, A., Dimbwadyo-Terrer, I., Trincado-Alonso, F., Monasterio-Huelin, F., Torricelli, D., and Gil-Agudo, A. (2014). Quantitative assessment based on kinematic measures of functional impairments during upper extremity movements: A review. *Clinical Biomechanics*, 29(7):719–727.

Debevec, P. (2008). Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *ACM SIGGRAPH*, page 32. ACM.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

Deniz, O., Serrano, I., Bueno, G., and Kim, T.-K. (2014). Fast violence detection in video. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, volume 2, pages 478–485. IEEE.

Dickinson, S. J., Pentland, A. P., and Rosenfeld, A. (1990). Qualitative 3-d shape reconstruction using distributed aspect graph matching. In *Computer Vision, 1990. Proceedings, Third International Conference on*, pages 257–262. IEEE.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Dryden, I. L. and Mardia, K. (1998). *Statistical shape analysis*, volume 4. John Willey & Sons.

Duan, X., Sun, H., and Peng, L. (2013). Riemannian means on special euclidean group and unipotent matrices group. *The Scientific World Journal*, 2013.

Endres, D., Meirovitch, Y., Flash, T., and Giese, M. A. (2013). Segmenting sign language into motor primitives with bayesian binning. *Frontiers in computational neuroscience*, 7.

Escorcia, V. and Niebles, J. C. (2013). Spatio-temporal human-object interactions for action recognition in videos. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 508–514.

Fabian Caba Heilbron, Victor Escorcia, B. G. and Niebles, J. C. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970.

Fanello, S., Gori, I., and Pirri, F. (2010). Arm-hand behaviours modelling: From attention to imitation. In *Advances in Visual Computing*, pages 616–627.

Favaro, P. and Soatto, S. (2005). A geometric approach to shape from defocus. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):406–417.

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941.

Ferguson, T. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent advances in Statist.*, 1:287–302.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *Ann. Stat.*, pages 209–230.

Filip, J. and Vávra, R. (2014). Template-based sampling of anisotropic BRDFs. *Comp. Graph. Forum*.

Flaherty, F. and do Carmo, M. (2013). *Riemannian Geometry*. Mathematics: Theory & Applications. Birkhäuser Boston.

Flash, T. and Handzel, A. A. (2007). Affine differential geometry analysis of human arm movements. *Bio. Cyb.*, 96(6):577–601.

Flash, T. and Hochner, B. (2005). Motor primitives in vertebrates and invertebrates. *Curr. Op. in Neurob.*, 15(6):660–666.

Flash, T., Meirovitch, Y., and Barliya, A. (2013). Models of human movement: Trajectory planning and inverse kinematics studies. *RAS*, 61(4):330–339.

Fletcher, P., Lu, C., Pizer, S., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. on Medical Imaging*.

Gams, A., Petrič, T., Do, M., Nemec, B., Morimoto, J., Asfour, T., and Ude, A. (2016). Adaptation and coaching of periodic motion primitives through physical and visual interaction. *RAS*, 75:340–351.

Gao, Y., Liu, H., Sun, X., Wang, C., and Liu, Y. (2016). Violence detection using oriented violent flows. *Image and vision computing*, 48:37–41.

Gates, D. H., Walters, L. S., Cowley, J., Wilken, J. M., and Resnik, L. (2016). Range of motion requirements for upper-limb activities of daily living. *American J. of Occupational Therapy*, 70(1).

Ghanem, B., Niebles, J. C., Snoek, C., Heilbron, F. C., Alwassel, H., Khrisna, R., Escorcia, V., Hata, K., and Buch, S. (2017). Activitynet challenge 2017 summary. *arXiv:1710.08011*.

Gong, D. and Medioni, G. (2011). Dynamic manifold warping for view invariant action recognition. In *ICCV, 2011 IEEE International Conference on*. IEEE.

Gong, D., Medioni, G., and Zhao, X. (2014). Structured time series analysis for human action segmentation and recognition. *TPAMI*, 36(7):1414–1427.

Gorür, D. (2007). *Nonparametric Bayesian Discrete Latent Variable Models for Unsupervised Learning*. PhD thesis, Max Planck Institute for Biological Cybernetics.

Gracia, I. S., Suarez, O. D., Garcia, G. B., and Kim, T.-K. (2015). Fast fight detection. *PloS one*, 10(4):e0120448.

Grosse, R., Johnson, M., Adelson, E. H., and Freeman, W. (2009). Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, pages 2335–2342.

Hamill, J. and Knutzen, K. M. (2006). *Biomechanical basis of human movement*. Lippincott Williams & Wilkins.

Harandi, M. T., Salzmann, M., and Hartley, R. (2014). From manifold to manifold: geometry-aware dimensionality reduction for spd matrices. In *ECCV 2014*. Springer.

Hartley, R., Trumpf, J., Dai, Y., and Li, H. (2013). Rotation averaging. *International journal of computer vision*, 103(3):267–305.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., and Davis, L. S. (2016). Learning temporal regularity in video sequences. In *CVPR*.

Hassner, T., Itcher, Y., and Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 1–6. IEEE.

Hen, Y. W. and Paramesran, R. (2009). Single camera 3d human pose estimation: A review of current techniques. In *Proc. of the Int'l Conf. for Technical Postgraduates (TECHPOS)*, pages 1–8. IEEE.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Hogan, N. and Sternad, D. (2012). Dynamic primitives of motor behavior. *Biological cybernetics*, pages 1–13.

Hoiem, D., Efros, A. A., and Hebert, M. (2005). Automatic photo pop-up. *ACM Transactions on Graphics (TOG)*, 24(3):577–584.

Holte, M. B., Moeslund, T. B., and Fihl, P. (2010). View-invariant gesture recognition using 3D optical flow and harmonic motion context. *Comp. Vis. and Im. Underst.*, 114(12):1353–1361.

Horn, B. K. (1977). Understanding image intensities. *Artificial intelligence*, 8(2):201–231.

Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., and Schaal, S. (2013). Dynamical movement primitives: learning attractor models for motor behaviors. *Neural computation*, 25(2):328–373.

Inamura, T., Toshima, I., Tanie, H., and Nakamura, Y. (2004). Embodied symbol emergence based on mimesis theory. *Int. J. of Robotics Research*.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human 3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339.

Jain, S. and Neal, R. M. (2004). A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*.

Jiang, Y. G., Li, Z., and Chang, S. F. (2011). Modeling scene and object contexts for human action retrieval with few examples. *TCSVT*, pages 674–681.

Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2018a). End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2018b). End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X., et al. (2016). T-cnn: Tubelets with convolutional neural networks for object detection from videos. *arXiv preprint arXiv:1604.02532*.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Kendall, W. S. (1990). Probability, convexity, and harmonic maps with small image i: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 3(2):371–406.

Kim, H. J., Xu, J., Vemuri, B. C., and Singh, V. (2015). Manifold-valued dirichlet processes. In *Proceedings of the 2015 International Conference on Machine Learning. International Conference on Machine Learning*, volume 2015, pages 1199–1208.

Kober, J. and Peters, J. R. (2009). Policy search for motor primitives in robotics. In *Adv. in neural inf. proc. systems*, pages 849–856.

Koenderink, J. (1984). What does the occluding contour tell us about solid shape? *Perception*, 13(3):321–330.

Koenderink, J. (1990). *Solid Shape*. MIT.

Kooij, J. F., Liem, M., Krijnders, J. D., Andringa, T. C., and Gavrila, D. M. (2016). Multi-modal human aggression detection. *Computer Vision and Image Understanding*, 144:106–120.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). HMDB: a large video database for human motion recognition. In *ICCV*.

Kulić, D., Ott, C., Lee, D., Ishikawa, J., and Nakamura, Y. (2012). Incremental learning of full body motion primitives and their sequencing through human motion observation. *The Int. J. of Robotics Research*, 31(3):330–345.

Lacquaniti, F., Terzuolo, C., and Viviani, P. (1983). The law relating the kinematic and figural aspects of drawing movements. *Acta Psychologica*, 54(1–3).

Land, E. H. and McCann, J. (1971). Lightness and Retinex theory. *JOSA*, 61(1):1–11.

Lea, C., Reiter, A., Vidal, R., and Hager, G. D. (2016). Segmental spatiotemporal cnns for fine-grained action segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer.

Lehrmann, A. M., Gehler, P. V., and Nowozin, S. (2013). A non-parametric bayesian network prior of human pose. In *Proc. of the IEEE Int'l Conf. on Computer Vision*.

Levi, Z. and Gotsman, C. (2013). ArtiSketch: a system for articulated sketch modeling. *Comput. Graph. Forum*, 32(2):235–244.

Li, S. and Chan, A. B. (2014). 3d human pose estimation from monocular images with deep convolutional neural network. In *Proc. of the Asian Conf. on Computer Vision*, pages 332–347. Springer.

Li, Y., Fermuller, C., Aloimonos, Y., and Ji, H. (2010). Learning shift-invariant sparse representation of actions. In *CVPR*, pages 2630–2637.

Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., and Snoek, C. G. (2018). Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50.

Liang, J., Park, F., and Zhao, H. (2013). Robust and efficient implicit surface reconstruction for point clouds based on convexified image segmentation. *Journal of Scientific Computing, Volume 54 Issue 2-3*.

Liebowitz, D., Criminisi, A., and Zisserman, A. (1999). Creating architectural models from images. In *Computer Graphics Forum*, volume 18, pages 39–50. Wiley Online Library.

Lillo, I., Niebles, J. C., and Soto, A. (2016). A hierarchical pose-based approach to complex action understanding using dictionaries of actionlets and motion poselets. In *CVPR*.

Lin, H.-T., Lin, C.-J., and Weng, R. C. (2007). A note on platt's probabilistic outputs for support vector machines. *Machine learning*, 68(3):267–276.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

Liu, Z., Zhu, J., Bu, J., and Chen, C. (2015). A survey of human pose estimation: The body parts parsing based methods. *J. of Visual Communication and Image Representation*, 32:10–19.

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimate. *The Ann. Statist.*, 12:351–357.

Loper, M., Mahmood, N., and Black, M. J. (2014). Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248.

Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. In *ACM SIGGRAPH*, volume 21, pages 163–169.

Lovell, D., Adams, R. P., and Mansingka, V. (2012). Parallel markov chain monte carlo for dirichlet process mixtures. In *Workshop on Big Learning, NIPS*.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proc. of the Int'l Conf. on Computer Vision*, volume 2, pages 1150–1157. IEEE.

Lu, C., Shi, J., and Jia, J. (2013). Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727.

Lu, J., Xu, R., and Corso, J. J. (2015). Human action segmentation with hierarchical supervoxel consistency. In *CVPR*, pages 3762–3771.

Luo, R. and Berenson, D. (2015). A framework for unsupervised online human reaching motion recognition and early prediction. In *IROS*, pages 2426–2433.

Lv, F. and Nevatia, R. (2006). Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *ECCV 2006*. Springer.

Mabrouk, A. B. and Zagrouba, E. (2018). Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications*, 91:480–491.

Magda, S., Kriegman, D. J., Zickler, T., and Belhumeur, P. N. (2001). Beyond Lambert: Reconstructing surfaces with arbitrary BRDFs. In *ICCV*, pages 391–398.

Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981. IEEE.

Malik, J. and Rosenholtz, R. (1997). Computing local surface orientation and shape from texture for curved surfaces. *International journal of computer vision*, 23(2):149–168.

Mallick, S. P., Zickler, T. E., Kriegman, D. J., and Belhumeur, P. N. (2005). Beyond Lambert: Reconstructing specular surfaces using color. In *CVPR*, pages 619–626.

Mandery, C., Terlemez, O., Do, M., Vahrenkamp, N., and Asfour, T. (2015). The kit whole-body human motion database. In *ICAR*, pages 329–336.

Maoz, U. and Flash, T. (2014). Spatial constant equi-affine speed and motion perception. *J. of Neurophysiology*, 111(2):336–349.

Marr, D. and Vaina, L. (1982). Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London B: Biological Sciences*, 214(1197):501–524.

Matusik, W., Pfister, H., Brand, M., and McMillan, L. (2003). A data-driven reflectance model. In *ACM SIGGRAPH*, pages 759–769.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., and Theobalt, C. (2018). Single-shot multi-person 3d pose estimation from monocular rgb. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2):90–126.

Mohammadi, S., Kiani, H., Perina, A., and Murino, V. (2015). Violence detection in crowded scenes using substantial derivative. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE.

Mohammadi, S., Perina, A., Kiani, H., and Murino, V. (2016). Angry crowds: Detecting violent events in videos. In *European Conference on Computer Vision*, pages 3–18. Springer.

Mori, G. and Malik, J. (2006). Recovering 3d human body configurations using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(7):1052–1062.

Moro, F. L., Tsagarakis, N. G., and Caldwell, D. G. (2012). On the kinematic motion primitives (kmps)–theory and application. *Frontiers in neurorobotics*, 6.

Mousavi, H., Mohammadi, S., Perina, A., Chellali, R., and Mur, V. (2015). Analyzing tracklets for the detection of abnormal crowd behavior. In *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–155. IEEE.

Munro, P. and Zipser, D. (1989). Image compression by back propagation: an example of extensional programming. *Models of cognition: rev. of cognitive science*, 1:208.

Narihira, T., Maire, M., and Yu, S. X. (2015). Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*.

Natola, F., Ntouskos, V., and Pirri, F. (2015a). Collaborative activities understanding from 3d data. *Doctoral Consortium on Pattern Recognition Applications and Methods (DCPRAM)*.

Natola, F., Ntouskos, V., Pirri, F., and Sanzari, M. (2016). Single image object modeling based on brdf and r-surfaces learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4414–4423.

Natola, F., Ntouskos, V., Sanzari, M., and Pirri, F. (2015b). Bayesian non-parametric inference for manifold based mocap representation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4606–4614.

Nayar, S., Ikeuchi, K., and Kanade, T. (1991). Surface reflection: physical and geometrical perspectives. *TPAMI*, 13(7):611–634.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. of Comp. and Graph Stat.*, 9(2):249–265.

Nealen, A., Igarashi, T., Sorkine, O., and Alexa, M. (2007). FiberMesh. *ACM TOG*, 26(3):41.

Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Le, Q. V., and Ng, A. Y. (2011). On optimization methods for deep learning. In *ICML*, pages 265–272.

Nicodemus, F. E. (1965). Directional reflectance and emissivity of an opaque surface. *Applied optics*, 4(7):767–775.

Nievas, E. B., Suarez, O. D., García, G. B., and Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In *International conference on Computer analysis of images and patterns*, pages 332–339. Springer.

Ntouskos, V., Papadakis, P., and Pirri, F. (2013). Discriminative sequence back-constrained gp-lvm for mocap based action recognition. In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 87–96.

Ntouskos, V., Papadakis, P., and Pirri, F. (2015a). Probabilistic discriminative dimensionality reduction for pose-based action recognition. In *Pattern Recognition Applications and Methods*, volume 318 of *Advances in Intelligent Systems and Computing*, pages 137–152.

Ntouskos, V., Sanzari, M., Cafaro, B., Nardi, F., Natola, F., Pirri, F., and Ruiz, M. (2015b). Component-wise modeling of articulated objects. *International Conference on Computer Vision (ICCV)*.

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. (2012). Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. In *CVPRW, 2012 IEEE Computer Society Conference on*.

Oliva, A. and Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11:520–527.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325.

Osher, S. and Fedkiw, R. (2003). *Level Set Methods and Dynamic Implicit Surfaces*. Springer.

Oswald, M., Eno, T., and Cremers, D. (2012). Fast and globally optimal single view reconstruction of curved objects. In *CVPR*, pages 534–541.

Oswald, M., Töppe, E., Nieuwenhuis, C., and Cremers, D. (2013). A review of geometry recovery from a single image focusing on curved object reconstruction. In *Innovations for shape analysis, models and algorithms*, pages 343–378. Springer.

Ouyang, W., Chu, X., and Wang, X. (2014). Multi-source deep learning for human pose estimation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2329–2336.

Oxholm, G. and Nishino, K. (2012). Shape and reflectance from natural illumination. In *ECCV*, pages 528–541. Springer.

Packer, B., Saenko, K., and Koller, D. (2012). A combined pose, object, and feature model for action understanding. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1378–1385.

Paisley, J., Wang, C., Blei, D. M., and Jordan, M. I. (2015). Nested hierarchical dirichlet processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):256–270.

Papadhimitri, T. and Favaro, P. (2013). A new perspective on uncalibrated photometric stereo. In *CVPR*, pages 1474–1481.

Park, D.-H., Hoffmann, H., Pastor, P., and Schaal, S. (2008). Movement reproduction and obstacle avoidance with dynamic movement primitives and potential fields. In *ICHR*, pages 91–98.

Pastor, P., Hoffmann, H., Asfour, T., and Schaal, S. (2009). Learning and generalization of motor skills by learning from demonstration. In *ICRA*, pages 763–768.

Pavllo, D., Feichtenhofer, C., Grangier, D., and Auli, M. (2019). 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pentland, A. P. (1986). Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28(3):293–331.

Pirri, F. and Pizzoli, M. (2011). *Knowing, Reasoning, and Acting Essays in Honour of Hector J. Levesque*, chapter Inference about Actions: Levesque's view on action ability and Dirichlet processes.

Pirsiavash, H. and Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854.

Pitman, J. (2006). *Combinatorial Stochastic Processes: Ecole D'Eté de Probabilités de Saint-Flour XXXII-2002*. Springer.

Plagemann, C., Ganapathi, V., Koller, D., and Thrun, S. (2010). Real-time identification and localization of body parts from depth images. In *2010 IEEE International Conference on Robotics and Automation*, pages 3108–3113.

Plantinga, S. and Vegter, G. (2006). Computing contour generators of evolving implicit surfaces. *ACM TOG*, 25(4):1243–1280.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Polyakov, F. (2017). Affine differential geometry and smoothness maximization as tools for identifying geometric movement primitives. *Biological cybernetics*, 111(1):5–24.

Pons-Moll, G., Fleet, D., and Rosenhahn, B. (2014). Posebits for monocular human pose estimation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2337–2344.

Poppe, R. (2007). Vision-based human motion analysis: An overview. *Computer vision and Image Understanding*, 108(1):4–18.

Prasad, M. and Fitzgibbon, A. (2006). Single view reconstruction of curved surfaces. In *CVPR*, pages 1345–1354.

Prasad, M., Fitzgibbon, A., Zisserman, A., and Van Gool, L. (2010). Finding nemo: Deformable object class modelling using curve matching. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1720–1727. IEEE.

Prasad, M., Zisserman, A., and Fitzgibbon, A. (2006). Single view reconstruction of curved surfaces. In *CVPR*, pages 1345–1354.

Prest, A., Ferrari, V., and Schmid, C. (2013). Explicit modeling of human-object interactions in realistic videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):835–848.

Prest, A., Schmid, C., and Ferrari, V. (2012). Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):601–614.

Ramanathan, V., Li, C., Deng, J., Han, W., Li, Z., Gu, K., Song, Y., Bengio, S., Rosenberg, C., and Fei-Fei, L. (2015). Learning semantic relationships for better action retrieval in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1100–1109.

Reddy, D., Agrawal, A., and Chellappa, R. (2009). Enforcing integrability by error correction using l1-minimization. In *CVPR*, pages 2350–2357.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Richter, S. R. and Roth, S. (2015). Discriminative shape from shading in uncalibrated illumination. In *CVPR*, pages 1128–1136.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested dirichlet process. *Journal of the American Statistical Association*.

Romeiro, F. and Zickler, T. (2010). Inferring reflectance under real-world illumination. Technical report, Cambridge, MA.

Rosenfeld, A. and Ullman, S. (2016). Action classification via concepts and attributes. *arXiv preprint arXiv:1605.07824*.

Sanzari, M., Natola, F., Nardi, F., Ntouskos, V., Qudseya, M., and Pirri, F. (2015). Rigid toll affordance matching points of regard. *IROS Workshop: Learning Object Affordances: a fundamental step to allow prediction, planning and tool use?*

Sanzari, M., Ntouskos, V., and Pirri, F. (2016). Bayesian image based 3d pose estimation. *European Conference on Computer Vision (ECCV)*.

Sanzari, M., Ntouskos, V., and Pirri, F. (2019). Discovery and recognition of motion primitives in human activities. *PloS one*, 14(4):e0214499.

Saxena, A., Sun, M., and Ng, A. Y. (2009). Make3d: Learning 3d scene structure from a single still image. *TPAMI*, 31(5):824–840.

Schmidt, M., Kim, D., and Sra, S. (2012). Projected newton-type methods in machine learning. *Optimization for Machine Learning*, page 305.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Sigal, L., Balan, A. O., and Black, M. J. (2009). HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *IJCV*, 87(1):4.

Sigal, L. and Black, M. J. (2006). Predicting 3d people from 2d pictures. In *Articulated Motion and Deformable Objects*, pages 185–195. Springer.

Simo-Serra, E., Ramisa, A., Alenyà, G., Torras, C., and Moreno-Noguer, F. (2012). Single image 3d human pose estimation from noisy observations. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2673–2680. IEEE.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576.

Soomro, K., Zamir, A., and Shah, M. (2012). Ucf101-action recognition data set.

Strang, G. and Fix, G. J. (1973). *An analysis of the finite element method*, volume 212. Prentice-Hall.

Straub, J., Chang, J., Freifeld, O., and Fisher III, J. W. (2015). A dirichlet process mixture model for spherical data. In *AISTATS*.

Sudderth, E. B. (2006). *Graphical models for visual object recognition and tracking*. PhD thesis, MIT.

Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. *Center for Research in Computer Vision (CRCV), University of Central Florida (UCF)*.

Taylor, C. J. (2000). Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 677–684. IEEE.

Taylor, J., Shotton, J., Sharp, T., and Fitzgibbon, A. (2012). The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 103–110. IEEE.

Teh, Y. W. (2011). Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2012). Hierarchical dirichlet processes. *Journal of the american statistical association*.

Tekin, B., Rozantsev, A., Lepetit, V., and Fua, P. (2016). Direct prediction of 3d body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–1000.

Tekin, B., Sun, X., Wang, X., Lepetit, V., and Fua, P. (2015). Predicting people's 3d poses from short sequences. *arXiv preprint arXiv:1504.08200*.

Terzopoulos, D., Platt, J., Barr, A., and Fleischer, K. (1987). Elastically deformable models. In *ACM SIGGRAPH*, pages 205–214.

Terzopoulos, D., Witkin, A., and Kass, M. (1988a). Constraints on deformable models: Recovering 3d shape and nonrigid motion. *Artificial intelligence*, 36(1):91–123.

Terzopoulos, D., Witkin, A., and Kass, M. (1988b). Symmetry-seeking models and 3d object reconstruction. *International Journal of Computer Vision*, 1(3):211–221.

Thurau, C. and Hlaváč, V. (2007). n-grams of action primitives for recognizing human behavior. In *International Conference on Computer Analysis of Images and Patterns*, pages 93–100. Springer.

Thurau, C. and Hlavác, V. (2008). Pose primitive based human action recognition in videos or still images. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Ting, L. H., Chiel, H. J., Trumbower, R. D., Allen, J. L., McKay, J. L., Hackney, M. E., and Kesar, T. M. (2015). Neuromechanical principles underlying movement modularity and their implications for rehabilitation. *Neuron*, 86(1):38–54.

Tome, D., Russell, C., and Agapito, L. (2017). Lifting from the deep: Convolutional 3d pose estimation from a single image. *CVPR 2017 Proceedings*, pages 2500–2509.

Tompson, J. J., Jain, A., LeCun, Y., and Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 1799–1807.

Töppe, E., Nieuwenhuis, C., and Cremers, D. (2013). Relative volume constraints for single view 3D reconstruction. In *CVPR*, pages 177–184.

Töppe, E., Oswald, M. R., Cremers, D., and Rother, C. (2011). Image-based 3d modeling via cheeger sets. In *Computer Vision–ACCV 2010*, pages 53–64. Springer.

Toshev, A. and Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1653–1660.

Turaga, P., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: A survey. *Trans. on Circuits and Systems for Video Technology*, 18(11):1473–1488.

Ureche, A. L. P., Umezawa, K., Nakamura, Y., and Billard, A. (2015). Task parameterization using continuous constraints extracted from human demonstrations. *IEEE Trans. Robot.*

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Varol, G., Laptev, I., and Schmid, C. (2018). Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517.

Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., and Schmid, C. (2017). Learning from synthetic humans. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*.

Vasilyev, Y., Adato, Y., Zickler, T., and Ben-Shahar, O. (2008). Dense specular shape from multiple specular flows. In *CVPR*, pages 1–8.

Vecchio, D. D., Murray, R. M., and Perona, P. (2003). Decomposition of human motion into dynamics-based primitives with application to drawing tasks. *Automatica*, 39(12):2085–2098.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vese, L. A. and Chan, T. F. (2002). A multiphase level set framework for image segmentation using the mumford and shah model. *IJCV*, 50(3):271–293.

Vicente, S. and Agapito, L. (2013). Balloon shapes: reconstructing and deforming objects with volume from images. In *3DV*, pages 223–230.

Vicente, S., Carreira, J., Agapito, L., and Batista, J. (2014). Reconstructing pascal voc. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 41–48. IEEE.

Viviani, P. and Flash, T. (1995). Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning. *J. of Exp. Psy.: Human Perception and Performance*, 21(1):32.

Viviani, P. and Schneider, R. (1991). A developmental study of the relationship between geometry and kinematics in drawing movements. *J. of Experimental Psychology: Human Perception and Performance*, 17(1).

Wang, C., Wang, Y., Lin, Z., Yuille, A., and Gao, W. (2014a). Robust estimation of 3d human poses from a single image. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2361–2368.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2014b). Learning actionlet ensemble for 3d human action recognition. *PAMI, IEEE Transactions on*.

Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., and Ogunbona, P. O. (2016). Action recognition from depth maps using deep convolutional neural networks. *IEEE Transactions on Human-Machine Systems*, 46(4):498–509.

Weinland, D., Ronfard, R., and Boyer, E. (2006). Automatic discovery of action taxonomies from multiple views. In *CVPR*, volume 2, pages 1639–1645.

Weinstock, R. (1974). *Calculus of variations: with applications to physics and engineering*. Courier Dover Publications.

West, M. (1992). *Hyperparameter estimation in Dirichlet process mixture models*. Duke University ISDS Discussion Paper# 92-A03.

Xiong, Y., Chakrabarti, A., Basri, R., Gortler, S. J., Jacobs, D. W., and Zickler, T. (2015). From shading to local shape. *TPAMI*, 37(1):67–79.

Xu, L., Gong, C., Yang, J., Wu, Q., and Yao, L. (2014). Violent video detection based on mosift feature and sparse coding. In *ICASSP*, pages 3538–3542.

Yang, G., Yin, Y., and Man, H. (2013a). Human object interactions recognition based on social network analysis. In *2013 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–4.

Yang, J., Nguyen, M. N., San, P. P., Li, X. L., and Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yang, Y. and Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890.

Yang, Y., Saleemi, I., and Shah, M. (2013b). Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE transactions on pattern analysis and machine intelligence*, 35(7).

Yao, B. and Fei-Fei, L. (2010a). 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.

Yao, B. and Fei-Fei, L. (2010b). Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24.

Yasin, H., Iqbal, U., Krüger, B., Weber, A., and Gall, J. (2015). 3d pose estimation from a single monocular image. *arXiv preprint arXiv:1509.06720*.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702.

Zefran, M., Kumar, V., and Croke, C. (1998). On the generation of smooth three-dimensional rigid body motions. *IEEE Trans. on Robotics and Automation*.

Zeisl, B., Zach, C., and Pollefeys, M. (2014). Variational regularization and fusion of surface normal maps. In *3DV*, volume 1, pages 601–608.

Zhang, L., Dugas-Phocion, G., Samson, J.-S., and Seitz, S. M. (2002). Single-view modelling of free-form scenes. *The Journal of Visualization and Computer Animation*, 13(4):225–235.

Zhang, R., Tsai, P.-S., Cryer, J., and Shah, M. (1999). Shape-from-shading: a survey. *TPAMI*, 21(8):690–706.

Zhou, F. and De la Torre, F. (2014). Spatio-temporal matching for human detection in video. In *Proc. of the European Conf. on Computer Vision*, pages 62–77. Springer.

Zhou, P., Ding, Q., Luo, H., and Hou, X. (2018). Violence detection in surveillance video using low-level features. *PLoS one*, 13(10):e0203668.

Zhou, X., Zhu, M., Leonardos, S., Derpanis, K., and Daniilidis, K. (2016). Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proc of the IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE.