Extracting human characteristics from motion using machine learning : the case of identity in Sign Language

Félix Bigand

université
PARIS-SACLAY

Extracting human characteristics from motion using machine learning: the case of identity in Sign Language

Extraction de caractéristiques humaines dans le mouvement par apprentissage automatique : l'exemple de l'identité en Langue des Signes

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : informatique
Unité de recherche : Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France
Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 22 novembre 2021, par

Félix BIGAND

Composition du Jury

Bastien Berret Professeur des universités, CIAMS, Université Paris-Saclay, Institut Universitaire de France	Président
Hervé Abdi Full professor, School of Behavioral and Brain Sciences, The University of Texas at Dallas	Rapporteur & Examinateur
Frédéric Bevilacqua Directeur de recherche, STMS, IRCAM, CNRS, Sorbonne Université	Rapporteur & Examinateur
Marion Blondel Chargée de recherche, SFL, CNRS, Université Paris8	Examinatrice
John McDonald Associate professor, School of Computing, DePaul University, Chicago	Examinateur

Direction de la thèse

Annelies Braffort Directrice de recherche, CNRS, LISN, Université Paris-Saclay	Directrice de thèse
Elise Prigent Maîtresse de conférences, LISN, Université Paris-Saclay	Co-encadrante de thèse

Thèse de doctorat
NNT: 2021UPASG090

"*Hélas ! une foule d'hommes, tous distingués par l'ampleur de la boîte cérébrale et par la lourdeur, par les circonvolutions de leur cervelle ; des mécaniciens, des géomètres enfin ont déduit des milliers de théorèmes, de propositions, de lemmes, de corollaires sur le mouvement appliqué aux choses, ont révélé les lois du mouvement céleste, ont saisi les marées dans tous leurs caprices et les ont enchaînées dans quelques formules d'une incontestable sécurité marine; mais personne, ni physiologiste, ni médecin sans malades, ni savant désœuvré, ni fou de Bicêtre, ni statisticien fatigué de compter ses grains de blé, ni quoi que ce soit d'humain, n'a voulu penser aux lois du mouvement appliqué à l'homme !*"

Honoré de Balzac
Théorie de la démarche (1833)

Abstract

Despite the large number of deaf individuals throughout the world using Sign Languages (SLs) to communicate, the vast majority of current communication tools only rely on spoken or written languages. Many technological barriers must be tackled in order to provide communication tools in SLs in the same way as for spoken languages, in particular by developing SL generation models.

The progress of motion capture (mocap) systems has considerably improved SL generation models, allowing for the animation of highly natural and comprehensible virtual signers. It has raised, however, an unexpected problem related to the human ability to identify individuals from their movements. Compared to the auditory domain where a speaker can remain anonymous by modifying specific voice characteristics, little is known about the motion features that characterize a signer's identity. Producing anonymized content with virtual signers is a crucial problem (e.g., for anonymized testimonies on journalistic websites). Yet, current research about person identification from SL motion remains sparse (if any).

Identity can be extracted from human movements, such as walking or dancing. What about SL, whose movements are not only constrained by biomechanical rules, but also by linguistic ones? The present thesis aimed to (1) gain insights into how the complex structure of SL movements can be modeled; (2) assess whether human perceivers actually manage to identify signers from SL motion; (3) determine the motion features allowing for signer identification and (4) develop computational models to control identity in SL motion generation. On the basis of a 3D mocap corpus in French Sign Language, the present thesis provides five main contributions.

First, we investigated the spectral content of the mocap data of spontaneous LSF. This study revealed that SL motion can be limited to a 0–12-Hz bandwidth, which is substantially wider than state-of-the-art estimates on isolated signs. These results suggest that SL motion involves higher frequencies in real-life conditions.

Additionally, we used Principal Component Analysis (PCA) to decompose spontaneous SL discourses into elementary movements called principal movements (PMs). Although the mocap data were not synchronized in time across signers and examples, PMs extracted separately for each signer and PMs extracted from the data of all signers were highly similar and explained the same amount of variance. These results suggest that SL may have a common structure that can be decomposed into simpler patterns using PCA.

Using Point-Light Displays, a visual perception study then revealed that deaf perceivers managed to identify familiar signers above chance level. Combining computational analyses of both the mocap data and the participants responses, the results of this study revealed that mocap data contain sufficient information to identify signers, beyond simple cues related to morphology.

A machine learning model was then trained for the automatic identification of signers, from statistics of the mocap data. The performance of the model was not affected when information about size and shape of the signers was normalized and it remained (although it decreased) over five times superior to chance level when posture normalization was applied. These findings demonstrate that a signer's identity

can be characterized by specific statistics of kinematic features, beyond information related to size, shape and posture.

Finally, a synthesis algorithm is proposed in order to re-synthesize natural SL movements for which the identity of the signer is modified. To do so, the identity-specific feature statistics (extracted by the model above) are manipulated. For instance, the impact of identity-specific features of the signer can be reduced (i.e., anonymization), and the importance of identity-specific features of another signer can be increased (i.e., identity conversion).

Résumé en français

Malgré le grand nombre de personnes sourdes utilisant les langues des signes (LS) pour communiquer, la majorité des outils actuels de communication ne reposent que sur les langues parlées ou écrites. De nombreux obstacles technologiques doivent être surmontés afin d'outiller les LS de la même manière que les langues parlées, en particulier en développant des modèles de génération des LS.

Les progrès des systèmes de capture de mouvement (mocap) ont considérablement amélioré les modèles de génération des LS, permettant d'animer des signeurs virtuels réalistes et compréhensibles. Ils ont cependant soulevé un problème inattendu, celui de l'identification du signeur à partir de ses mouvements. Comparé au domaine auditif où un locuteur peut rester anonyme si l'on modifie certains aspects de sa voix, on ne sait que peu de choses sur les aspects du mouvement qui permettent d'identifier un signeur.

L'identité peut être extraite de mouvements humains, comme la marche ou la danse. Qu'en est-il de la LS, dont les mouvements ne sont pas seulement contraints par des règles biomécaniques, mais également par des règles linguistiques ? Cette thèse vise à (1) comprendre comment la structure complexe des mouvements de la LS peut être modélisée ; (2) évaluer si les humains parviennent à identifier des signeurs à partir de leurs mouvements ; (3) déterminer les aspects du mouvement permettant l'identification du signeur et (4) développer des modèles informatiques pour contrôler l'identité lors de la génération du mouvement des LS. Sur la base d'un corpus de mocap 3D en Langue des Signes Française, cette thèse fournit cinq contributions principales.

Premièrement, nous avons étudié le contenu spectral des données mocap issues de discours spontanés en LSF. Selon cette étude, le mouvement de la LSF peut être limité à une bande passante de 0-12 Hz, ce qui est considérablement plus large que les estimations antérieures réalisées sur des signes isolés. Ces résultats suggèrent que le mouvement de la LS implique des fréquences plus élevées en conditions réelles.

Nous avons également utilisé l'analyse en composantes principales (PCA) pour décomposer des discours spontanés de LS en mouvements principaux (PM). Les PM extraits séparément pour chaque signeur et ceux extraits pour l'ensemble des six signeurs sont très similaires, et expliquent la même quantité de variance. Ces résultats suggèrent que le LS doit avoir une structure commune, qui peut être décomposée en éléments simples à l'aide de la PCA.

Une étude de perception visuelle a ensuite évalué la capacité de participants sourds à identifier des signeurs à partir de stimuli « Point-Light ». En combinant des analyses informatiques des données mocap et des réponses des participants, les résultats de cette étude ont révélé que les données mocap contiennent suffisamment d'information pour identifier les signeurs, au-delà des indices liés à la morphologie.

Nous avons ensuite entraîné un modèle d'apprentissage automatique pour l'identification de signeurs, à partir de statistiques des données mocap. La performance du modèle n'a pas été affectée lorsque les informations sur la taille et la forme des corps des signeurs ont été normalisées. Elle est restée plus de cinq fois supérieure

au niveau du hasard lors de la normalisation de la posture. Ces résultats démontrent que l'identité d'un signeur peut être caractérisée par des statistiques liées à la dynamique, au-delà des informations liées à la taille, la forme et la posture.

Enfin, un algorithme de synthèse est proposé afin de re-synthétiser des mouvements de LS tout en modifiant l'identité du signeur. Pour ce faire, les statistiques spécifiques liées à l'identité (cf. modèle ci-dessus) sont manipulées. Par exemple, l'impact des caractéristiques spécifiques à l'identité du signeur peut être réduit (i.e., anonymisation), et l'importance des caractéristiques spécifiques à l'identité d'un autre signeur peut être augmentée (i.e., conversion d'identité).

Acknowledgements

I'm writing these acknowledgements somewhere between Paris and Rome, coming back from the defence of these three wonderful years of research as a PhD student. First of all, I would like to thank my supervisors Annelies Braffort and Elise Prigent, without whom all this adventure would not have been possible. From our very first interview I knew that I would work with people of rare kindness and openness. I am deeply grateful to both of you for the time and energy you have devoted to my work, even (more) in the complex pandemic period, which made the second and third years of this PhD quite challenging. I am additionally grateful to Annelies for having introduced me to the fascinating world of sign languages. Finally, beyond our research interactions, I feel very lucky for having met you both, and I am looking forward to coming back to the lab as soon (and as much) as I can!

I also want to thank Bastien Berret, who was a constant inspiration to me and who allowed me to refine technical aspects of my studies so many times. I learned so much from the discussions we had, despite they were sometimes very short! I am therefore looking forward to pursuing these chats and exchanges, now from my new position at the IIT (here again you inspired me!).

Then, I would like to thank the people who contributed to this unique chapter of my life, in both my work and personal life. First of all, I acknowledge the LIMSI (now LISN) staff for providing us such a joyful and enriching environment to work and the M&TALS group for the repeated fruitful discussions and the good times, with a special thought for Michael (my Spritz magician), Valentin (my inspiring office neighbour) and Marion (*in bocca al luppo* for the defence!). I thank the researchers I have met in my first year of PhD whose (even short) words and feedbacks motivated my work, at a time when it was challenged by some doubts and interrogations: Dominique Boutet, John McDonald, Jean-François Jégo, Boris Dauriac, Jules Françoise.

In particular, I am grateful to Hervé Abdi, whose fascinating work has grounded the whole present manuscript, and Frédéric Bevilacqua, who followed and discussed this work from almost the beginning. Thanks for having reviewed this manuscript, which we all know is a great deal of work. Relatedly, I am grateful to Hervé, Fred, John and Marion Blondel for having examined my work and giving me the privilege to discuss it with them during the defence.

I have a thought for Sam Norman-Haignere, who made me take my first steps in the research field and who I consider, still today, as a caring guide in the field, beyond being a brilliant and inspiring researcher. I also acknowledge the consequent feedbacks of Barbara about my writings, sharing me tips on how to write top-level academic English (thanks to Amélia too, for the English corrections).

On a more personal note, I would like to thank my fabulous band Panache!, for the incredible concerts and enchanted escapes during week-ends and summer holidays. In particular, I feel grateful to Amélia, Pierre and Raoul for your patience and for the magic of keeping our planets aligned in every circumstances. I thank my family: Papa, I am so thankful for the constant chats and brain stimulation; Maman, I am so grateful for your energy and we both know that a story starting from *Le Cabanon*

could not have ended badly; Vico, my litte brother I am so proud of, I can't wait to follow your new music and work. I also sincerely think that nothing would have been the same without all the people surrounding me during these years: Valérie, Philippe, Marie, Cécile, Alex, Paul, Charles, Samuel, my grandparents, Antoine et Anne, Liora, Lucas, Yves, Florian, Alain et Isabelle, Madhuri, to name a few...

The final word is for Juliette, who knows the ins and outs of this thesis so well that she could have defended it in my place. Thank you for all, the happiness that you bring to my life gives me the strength to keep going and to flourish in this crazy adventure that is research.

Somewhere between Paris and Rome, November 2021

À mon grand-père, Philippe.

Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	**An**alysis **O**f **Va**riance
ASL	**A**merican **S**ign **L**anguage
BSL	**B**ritish **S**ign **L**anguage
CoM	**C**enter **o**f **M**ass
DGS	German Sign Language (***D**eutsche **G**ebärden**s**prache*)
EV	**E**igen**V**ector
fps	**f**rames **p**er **s**econd
GRNN	**G**eneralized **R**egression **N**eural **N**etwork
HMM	**H**idden **M**arkov **M**odel
IMU	**I**nertial **M**easurement **U**nit
LSF	French Sign Language (***L**angue des **S**ignes **F**rançaise*)
M	**M**ean
ML	**M**achine **L**earning
Mocap	**Mo**tion **cap**ture
OA	medial knee **O**steo**A**rthritis
ORI	**Ori**ginal
PC	**P**rincipal **C**omponent
PCA	**P**rincipal **C**omponent **A**nalysis
PLD	**P**oint-**L**ight **D**isplay
PM	**P**rincipal **M**ovement
POST	**Post**ure-normalized
PSD	**P**ower **S**pectral **D**ensity
QoM	**Q**uantity **o**f **M**otion
RBF	**R**adial **B**asis **F**unction
RGB	**R**ed-**G**reen-**B**lue
RMS	**R**oot **M**ean **S**quare
SD	**S**tandard **D**eviation
SH	**Sh**ape-normalized
SI	**Si**ze-normalized
SL	**S**ign **L**anguage
SVM	**S**upport **V**ector **M**achine

Chapter 1

Introduction

The present thesis is the result of research in the fields of motion, computer science and visual perception. It investigates how complex human movements are structured, how they are perceived by human observers, and how human characteristics are encoded in motion patterns. In particular, this thesis tackles how identity information is encoded in the movements of signers in Sign Language.

1.1 The social impact of Sign Language automatic processing

Sign Languages (SLs) are the first languages of 70 million deaf people in the world (WFD, 2016). Like any other natural language, SLs follow specific rules defined by a linguistic system. In the case of SLs, this linguistic system is structured on a visual-gestural modality. SL users express themselves producing a continuous stream of movements with numerous body parts, such as hands, torso, or face. there are over 300 different SLs used around the world. They are distinct from spoken languages and have no written form. For instance in France, many deaf individuals have French Sign Language (LSF) as a first language, French being only a second one. Therefore, for deaf persons, reading written content means reading a second language, which is not always mastered (Holt, 1993).

Yet, the vast majority of existing communication tools are designed in spoken or written languages. Indeed, extensive research and developments have provided numerous tools for the automatic processing of spoken languages, including tasks such as speech recognition, speech segmentation or text-to-speech conversion. For a decade now, a wide variety of devices are equipped with voice assistants, such as Apple's Siri or Amazon's Alexa, which can interpret human speech and respond via synthesized voices. By contrast, although some promising SL applications have been developed in the past decade (e.g., *"Jade"*, *"Keia"* for LSF), further work is still needed to provide tools for SLs in the same way as for spoken languages. Improvements in SL automatic processing (that is SL automatic recognition, generation and translation) would thus allow breaking communication barriers faced by deaf SL users with most existing tools designed in spoken or written languages.

One reason for the current limitations of SL automatic processing is that SLs are poorly endowed and not yet fully described. Research on SL is recent and remains sparse, compared with that of spoken languages. Moreover, most SL research has focused on linguistics while the design of new SL technologies can benefit from other fields, such as computer science. For instance, developing computational models for the generation of SL movements could allow personal assistants to respond not only with synthesized voices, but also with synthesized signed movements. In that respect, there has been recent interesting developments toward the automatic production of SL messages via virtual signers (or signing avatars) (Filhol et al., 2017; Filhol and Mcdonald, 2020; Wolfe et al., 2011).

1.2 Improving virtual signers: toward natural motion

Virtual signers have many advantages. Unlike pre-recorded videos, virtual signers can be used in dynamic and interactive scenarios, as the content of the animation can be modified (Kipp et al., 2011). Moreover, the appearance of virtual signers can be manipulated, which allows adapting the agent to specific audiences. With a controlled appearance, virtual signers also provide new potentials for the production of anonymized SL messages, which was not possible with pre-recorded videos. Note that in the following sections, the main contributions of the present thesis will outline that controlling appearance, however, contributes to only some aspects of anonymization.

Despite the numerous advantages mentioned above, the use of virtual signers is still limited to date. The high number of body segments and the variety of linguistic structures involved in SL make it challenging to design efficient computational models that generate natural motion. This limitation can severely affect the perception of virtual signers, as the human brain is extremely sensitive to biological motion, that is the movements of humans and other vertebrates. Humans are able to detect actions (Johansson, 1973) as well as to derive information about other individuals from motion, such as gender (Cutting et al., 1978) or emotion (Venture et al., 2014). The ability to distinguish the dynamic regularities of biological motion from non-biological motion seems to appear very early in human life (Méary et al., 2007). Generation models thus have to ensure that virtual signers produce natural movements (i.e., that can be perceived as biological), in order to make them perceptually acceptable and comprehensible.

We are still far from having enough knowledge about SL motion perception to produce natural movements with purely synthetic animations. One way to overcome this limitation is to replay movements that were recorded on humans. Using motion capture (mocap) systems, the movements of real persons can be recorded with high accuracy and the virtual signers can be animated using the pre-recorded movements (Lu and Huenerfauth, 2010; Gibet, 2018). In addition to naturality, mocap also allows for the production of SL messages with high comprehensibility, which is another crucial challenge for the animation of virtual signers (Kipp et al., 2011).

1.3 The *naturality vs. anonymity* dilemma

Mocap systems provide more natural and comprehensible motion. They raise, however, another unexpected problem, notably related to person identification. As for spoken languages in the auditory domain, where voice parameters inform about a speaker's identity, a signer's identity could be conveyed by his or her movements. This observation questions the possibility to produce anonymized, non-identifiable, content with virtual signers. This problem is crucial (e.g., for sharing anonymized testimony) given that SLs have no written form (see Section 1.1). Compared to the auditory domain where a speaker can remain anonymous by modifying specific voice characteristics, little is known about the motion features that characterize a signer's identity and how these features could be manipulated in SL animations.

Up to now, current research about person identification in SL motion remains sparse (if any). Compared to prior evidence provided for other human movements, further research is needed to gain insights into the signer identification problem in

SL, which in particular raises two major questions. First, visual perception experiments could evaluate to which extent humans actually manage to infer identity from SL motion displayed as Point-Light Displays, as previously shown for other human movements (Troje et al., 2005; Loula et al., 2005; Sevdalis and Keller, 2009; Bläsing and Sauzet, 2018). Second, computational approaches, including machine learning, could shed light on how identity is encoded in the movement patterns of the signers, similarly to prior work on the gender attributes of gait (Troje, 2002a), or on the identity-specific features of dance (Carlson et al., 2020). These two approaches will be used in the present thesis to tackle the signer identification problem, as well as other computational methods in order to better understand the general kinematic properties and complex structure of SL movements, beyond the question of identification.

1.4 Outline of the dissertation

For all these reasons, based on prior studies in human perception and computational analysis of motion (Part I), the present thesis investigates the encoding of identity in SL motion. We present contributions in the computational processing, analysis and decomposition of SL mocap data (Part II), and in the extraction of identity from SL motion, using computational and visual perception methods (Part III).

Part I aims to introduce the context and related literature which form the theoretical foundations of the research carried out in this thesis. We first explore the underpinnings of motion perception from gait to Sign Language, and discuss how identity can be inferred from movements by human observers (Chapter 2). Then, we outline the crucial contributions that mocap and computational approaches, including machine learning, can have for gaining insights into the complex structure of SL movements and into the encoding of human characteristics in SL motion. We elaborate on the technical advances of mocap systems, on the motion representations that can be derived from these systems and on their contribution to SL research (Chapter 3). Moreover, we present how mocap data allow for quantitative analyses of motion and we discuss how computational models could provide further insights into how to extract identity-specific features from motion and how to manipulate these features in novel synthesized movements (Chapter 4).

Within this theoretical framework, the main contributions of the present thesis are twofold. First, kinematic analyses of French Sign Language (LSF) 3D mocap data are reported (Part II). The original 3D mocap corpus used in all of the following studies is described, a new version of the corpus is proposed and novel mocap data processing tools are presented (Chapter 5). Inspired by other fields of signal processing, time-frequency analyses of mocap data are then proposed in order to determine the kinematic bandwidth of SL (Chapter 6)[1]. Moreover, Principal Component Analysis (PCA) is applied to mocap data in order to test the decomposition of complex and non-synchronized SL movements into simpler, elementary, movements (Chapter 7)[2]. Additionally in this chapter, we question to which extent these elementary movements are identity-specific and discuss the potential limits of temporal-based approaches for investigating the encoding of identity information in motion.

The second main contribution of this thesis concerns the extraction of identity features from motion, from both perceptual and computational perspectives (Part

[1]Chapter 6 is partly reproduced from Bigand et al. (2021b).
[2]Chapter 7 is partly reproduced from Bigand et al. (2021a).

III). We first address this problem with a visual perception experiment, which provides the first insights into the actual ability of deaf perceivers to identify signers from mocap data (Chapter 8)[3]. We further present and discuss a machine learning framework, which is aimed to determine which parts of motion information are responsible for signer identification (Chapter 9)[4]. Furthermore, we propose a synthesis algorithm in order to manipulate identity-specific kinematic features in SL mocap animations while preserving the semantic content (e.g., for anonymization) (Chapter 10). The contributions of this thesis are finally summarized and discussed toward future research.

[3]Chapter 8 is partly reproduced from Bigand et al. (2020).
[4]Chapter 9 is partly reproduced from Bigand et al. (2021c).

Part I

Background and related work

Chapter 2

Human perception of biological and Sign Language motion

Human observers exhibit an impressive ability to process the information contained in the movements of their conspecifics, whether for recognizing actions or for inferring higher-level information, such as emotion, gender or identity. We elaborate on this ability in human movements, such as walking or dancing (Section 2.1), as well as for SL motion (Section 2.2). More specifically, we discuss the human ability to identify individuals from motion and outline why further investigations are necessary to better understand this ability, in particular for SL motion (Section 2.3).

2.1 Human perception of biological motion

Human perception of biological motion has been a focus of study for decades. Visual perception studies have shown that humans can efficiently infer a rich amount of information about their conspecifics from their movements (Section 2.1.1). Different theories have been proposed to determine the mechanisms that account for this impressive ability (Section 2.1.2). Moreover, in addition to the understanding of others' movements, prior work has outlined the ability of human observers to infer human characteristics from biological motion, notably identity (Section 2.1.3). A few studies have also investigated which parts of the motion information allow for the identification (Section 2.1.4).

2.1.1 The impressive human visual sensitivity to biological movements

The perception of biological motion, that is the movements of humans and other vertebrates, has been studied using Point-Light Displays (PLDs). These displays were constructed by attaching lights to the major joints of moving persons (Johansson, 1973; Johansson, 1976). PLDs isolate information given by motion cues from information given by other characteristics, such as shape or other aspects of the agent's body. Using PLDs, the demonstrations of Johansson were twofold. First, human observers never interpreted the static set of dots as a human body, but they were able to identify human movements, such as walking, running or dancing, when the dots were put into motion (Johansson, 1973). Second, observers managed to identify the human movements rapidly, with exposure times as short as 200 ms (Johansson, 1976). Moreover, 100 ms were enough for 40% of the observers to perceive a human body in the moving dots, although specifying the type of movement was too hard.

Further studies then have outlined the human ability to recognize various categories of movements (Dittrich, 1993; Bertenthal and Pinto, 1994; Poizner et al., 1981) as well as to infer other information such as intention (Runeson and Frykholm, 1983),

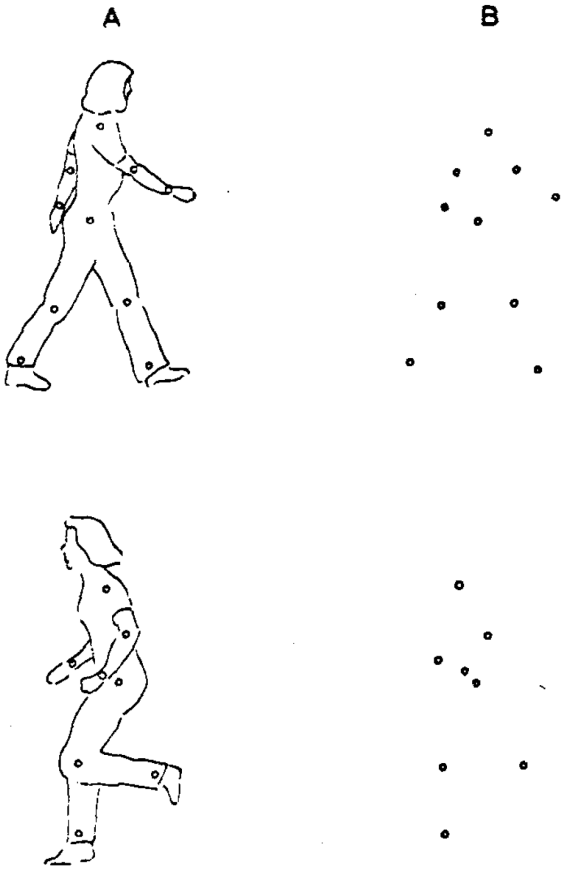


FIGURE 2.1: The first example of Point-Light Display (PLD) (Johansson, 1973). Schematic of the setup on walking or running persons (A) and the resulting PLD (B).

emotion (Brownlow et al., 1997; Dittrich et al., 1996; Venture et al., 2014; Atkinson et al., 2004; Camurri et al., 2003a; Lagerlöf and Djerf, 2009; Hietanen et al., 2004)) or musical expression (Luck et al., 2010; Dahl and Friberg, 2007; Davidson, 1993). Based on PLDs, humans are also able to rapidly detect specific information related to the task, such as the weight of a lifted box (Runeson and Frykholm, 1981) or the direction or speed of walkers (Jacobs and Shiffrar, 2005), even in greatly degraded conditions (e.g., adding masks to the PLDs, displaying a reduced number of body joints or drastically reducing exposure time) (Cutting et al., 1988; Neri et al., 1998; Jacobs et al., 2004).

Furthermore, this sensitivity to biological motion seems to appear very early in human life. Fox and McDaniel (1982) have demonstrated that infants 4 to 6 months of age preferred biological motion patterns (i.e., upright PLDs of walking persons) to artificially manipulated motion patterns (i.e., upside-down versions of the walkers' PLDs). This orientation-specific visual preference has also been reported for 3- an 5-month-old infants (Bertenthal et al., 1984; Bertenthal et al., 1987). Similarly to Johansson (1973), infants were able to discriminate between upright and inverted moving PLDs but not between static ones, which supports the idea that sensitivity to biological motion is developed early in human life. In addition to these studies using PLDs of multiple point lights, Méary et al. (2007) have also reported differences in the looking behavior of 4-day-old human neonates when perceiving biologicial or non-biological PLDs of one moving point light.

2.1.2 What's behind?

The human brain is extremely sensitive to biological movements. This raises the question of which mechanisms account for this sensitivity. In that respect, two classes of theories have been proposed. On the one hand, the visual perception of an action benefit from the observer's ability to produce the same action, that is motor experience. On the other hand, the visual sensitivity of human observers is enhanced by the elevated frequency with which they perceive others' movements in their environments, that is visual experience.

Studies in psychology (Prinz, 1997), neuroscience (Rizzolatti and Craighero, 2004) and computer science (Wolpert and Ghahramani, 2000) have supported the theory of an intrinsic linkage between visual and motor systems (i.e., perception-action coupling). Motion perception and production may share representations for the same actions (Prinz, 1997). Numerous evidence from neurophysiological data have supported motor simulation as having a crucial role in the perception of others' actions (Blakemore and Decety, 2001). In Rizzolatti et al. (2001), mirror neurons in the premotor cortex of the macaque monkey were found to respond both when performing an action and when observing the same action performed by another monkey. Imaging data of the human brain have revealed that comparable mechanisms can be located in the inferior frontal cortex, notably in Broca's area (Iacoboni et al., 1999). In addition, several psychological accounts have suggested that properties of the motor experience influences the visual perception of a person's own movements and of the movements of other people (Reed and Farah, 1995; Viviani and Stucchi, 1992; Grèzes et al., 2004; Knoblich and Flach, 2001; Jacobs and Shiffrar, 2005).

However, some characteristics of motion perception can hardly be explained by perception-action coupling. For instance, the perception of depth structure in PLDs is facilitated by usual viewpoints, which suggests that observers rely on visual experience (Bülthoff et al., 1998). According to Johansson (1973), the vividness with which observers perceive human movements may be due to previous experience with these movements. Neurophysiological data have also reported that brain response was modulated with visual experience (Grossman and Blake, 2001) in an area selective to biological motion (the posterior superior temporal sulcus (pSTS) (Grossman et al., 2000)).

The role of motor and visual experience in human perception of movements has been a central question in the field of visual and motion perception. Moreover, it has raised a further problem: person identification. Visual perception studies have suggested that observers perform better at identifying the moving person with their own moving PLDs than with those of their friends (Beardsworth and Buckner, 1981; Loula et al., 2005). In addition to this higher sensitivity to one own's movements, Jacobs et al. (2004) have shown that person identificiation from walking PLDs was more accurate when viewing possible rather than impossible gaits, which support perception-action coupling theories. In summary, although the several studies mentioned above have outlined the role of motor experience in various visual perception tasks on human movements, the specific problem of person identification is particularly suited to explore that question. Indeed, identifying one own's movements with higher accuracy than those of others support motor experience as key contributor to motion perception, as individuals hardly ever see their own movements, compared to those of others.

2.1.3 Person identification from human movements

Not only can human observers derive relevant information from others' movements, such as category of action, intention or emotion (see Section 2.1.1), but they can also determine the moving person's gender (Kozlowski and Cutting, 1977; Mather and Murdoch, 1994) and identity (Cutting and Kozlowski, 1977; Troje et al., 2005; Loula et al., 2005; Sevdalis and Keller, 2009; Stevenage et al., 1999). The first study by Cutting and Kozlowski (1977) has reported modest performance, but significantly above chance level, for human identification of walkers from PLDs. These results have suggested that, although the task was not easy, PLDs contained important information to allow for person identification. In Jacobs et al. (2004), observers were able to discriminate between the identity of two walkers with a 73% accuracy, when they previously had greatly interacted with each other (over 20 hours a week). Performance lowered to chance level (50 %) when participants had previously seen the walkers only 5 hours a week, or less. Troje et al. (2005) have confirmed that person identification from PLDs requires prior visual experience, or training, with the walkers' movements. After training steps to familiarize with the individuals to be recognized, identification performance of the participants reached 79%, over five times higher than chance level (14 %). These results have suggested that prior exposure to the others' movements (e.g., with prior social interaction or with pre-training) was required to accurately identify other people. However, beyond the human ability to recognize familiar individuals, Baragchizadeh et al. (2020) have recently demonstrated that motion cues also allow for the perceptual discrimination of the identity of unfamiliar people, without any prior training.

Using PLDs, behavioral studies have shown that the identity of individuals can be inferred from various human movements, such as walking (Cutting and Kozlowski, 1977; Troje et al., 2005; Jacobs et al., 2004; Stevenage et al., 1999) but also dancing (Bläsing and Sauzet, 2018) or clapping (Sevdalis and Keller, 2009). In particular, Loula et al. (2005) have investigated person identification from PLDs depicting a wide variety of different actions, such as jumping, hugging, boxing, running or ping-pong playing. Interestingly, the highest human performance in person identification occurred for PLDs of dancing, boxing, jumping and ping-pong playing individuals. Identification of walkers or runners was much lower, which is in line with the modest recognition reported by prior studies on point light walkers (Cutting and Kozlowski, 1977; Jacobs et al., 2004). These results suggest that person identification may be facilitated by movements with more complex spatiotemporal structures.

2.1.4 Unveiling the encoding of identity information in human movements: where do we stand?

Beyond the overall ability of humans to infer identity from motion, only a few visual perception studies have aimed to determine the cues that allow for the identification. Previous findings using PLDs have outlined that critical features for gender classification of gait seem to be in the frontal plane, which are, therefore, best visible in frontal view (Mather and Murdoch, 1994; Troje, 2002a). A similar behavioral study has also reported that human observers were better able to identify walkers from PLDs when presented in frontal view (Troje et al., 2005). However, although the frontal view allowed for an overall higher recognition, observers were better able to identify walkers in new viewpoints when they had been trained (i.e., familiarized with the walkers to be recognized) on half-profile PLDs. Moreover, no overall advantage for the frontal view has been reported by Westhoff and Troje (2007), whose

gait PLDs only included kinematic information. Therefore, although most of the critical information reported by prior studies seem to be in the frontal plane, half-profile and profile views may still provide critical information. According to Westhoff and Troje (2007), we could hypothesize a higher role of these viewpoints especially for kinematic information.

A few studies have assessed the role of specific motion components in identification. According to Troje et al. (2005), removing the walkers' size and shape information from PLDs had only low impact on human identification accuracy, which was still five to six times above chance level. More recently, Simhi and Yovel (2020) have conducted a virtual reality study with human participants, which highlighted that walking persons can be identified beyond face and body information. These results suggest that most of the information used for identification is conveyed by motion kinematics. The nature of such kinematic cues remains relatively unclear up to now. According to Troje et al. (2005) and Westhoff and Troje (2007), gait frequency may not play a major role in identification. The most critical information for identification seems to be conveyed by the first harmonic and the amplitude spectrum of walking patterns (Westhoff and Troje, 2007). Further investigation is needed to better understand the role of kinematic cues in the perception of an individual's identity, in particular for Sign Language (SL). One specific aspect of SL is to be governed not only by biomechanic rules, but also by linguistic ones, which may thus reveal SL-specific signatures for signers' identity.

2.2 Human perception of Sign Language motion

SLs are unique cases of biological motion, as not only are they constrained by biomechanical rules but also by linguistic ones. At the same time, SLs are unique languages as they involve a visual-gestural modality (Section 2.2.1). The information may thus be processed differently by human observers (in particular, signers (i.e., SL users)) when perceiving SL movements than when perceiving purely biological ones (Section 2.2.2).

In the following sections, although there is a wide variety of Sign Languages (SLs) throughout the world, we will sometimes refer to Sign Language (SL) in general when observations can be extended to all SLs.

2.2.1 Sign Language: language and gesture

SLs are languages that have naturally evolved in deaf communities throughout the world (Klima and Bellugi, 1979). Like spoken languages, SLs are governed by a linguistic system including syntactic, morphological and phonological structures (Emmorey, 2001). However, unlike spoken languages, the linguistic system of SL also includes other structures related to their visual-gestural modality. For instance, signers rely on an extensive use of space to build their SL discourse, as well as of iconicity, that is the strong iconic resemblance of the form of signs to what they represent (Sallandre and Cuxac, 2002). Note that there is no consensus on the description of SLs amongst linguists as yet. Furthermore, SLs are oral languages (i.e., involving face-to-face communication rather than written). A signer produces a message in SL and an observer perceives the message (Figure 2.2), like speakers and listeners in spoken languages.

Neurophysiological data have shown that, like speech production, SL execution activates Broca's area (Corina et al., 1999; Hickok et al., 1996), which suggests that

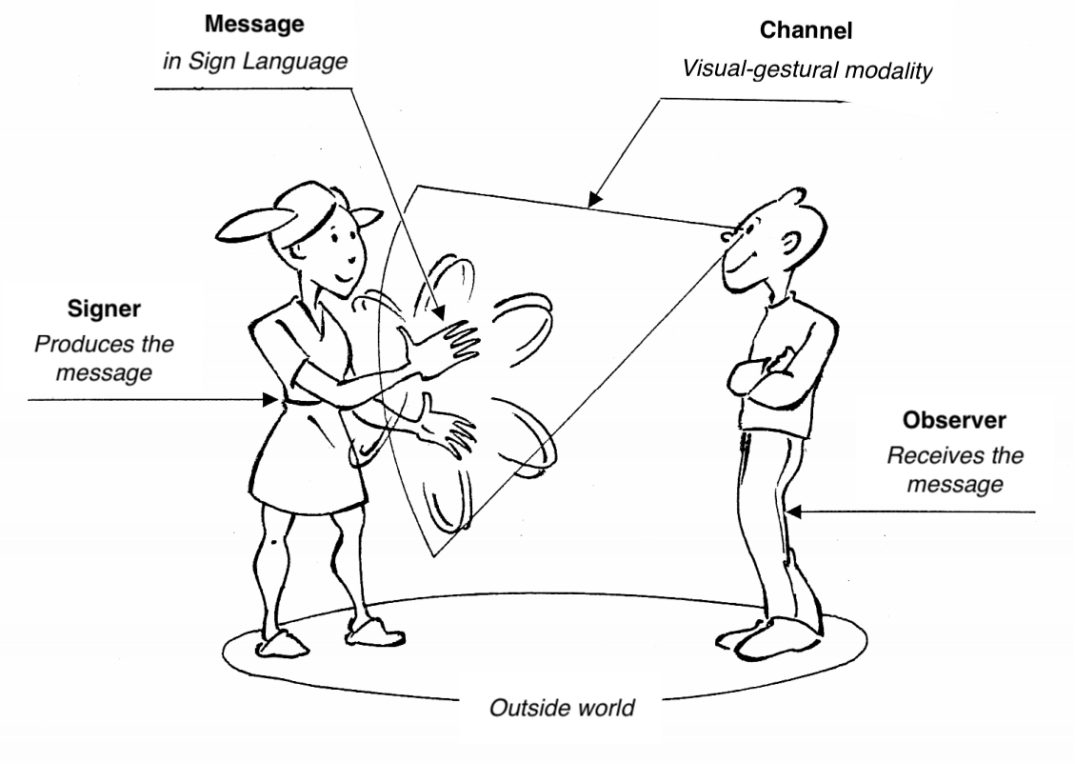


FIGURE 2.2: Sign Languages: oral languages via a visual-gestural modality (schematic taken from Guitteny (2006)).

the specificity of this brain region is not limited to speech processing. Neville et al. (1998) have reported that not only did American SL (ASL) elicited brain activation of native signers in areas involved in speech processing, but also in homologous areas within the right hemisphere. Furthermore, the right hemisphere has been shown to be involved in ASL processing only in native signers, by contrast with signers who had learnt ASL after puberty (Newman et al., 2002). This suggests that, on one hand SL shares common mechanisms with spoken languages, but on the other hand it has specific requirements that involve other brain regions.

Moreover, the reported co-activation of Broca's area during both the execution and the perception of SL by native signers (Okada et al., 2016) supports the perception-action coupling theory developed in biological movements (Section 2.1.2). As a reminder, neurophysiological studies have reported that Broca's area, which has been shown to be the motor area for speech, is involved in sensorimotor integration (Iacoboni et al., 1999). This suggests that human language may have evolved from a cortical system understanding movement, more specifically manual gestures (Corballis, 1999). SL is thus an intruiguing domain for the investigation of perception-action coupling, as it connects sensorimotor processes in language and gesture. Unlike other biological movements, such as walking or jumping, SL movements convey linguistic meaning and signers process SL movements of other signers in order to comprehend it. Similarly to the prior research on biological motion mentioned in Section 2.1, visual perception studies have allowed better understanding how signers perceive and process SL movements.

2.2.2 Visual perception studies on Sign Language motion

Following Johansson (1973), PLDs have been used to investigate SL perception, more specifically SL comprehension (Poizner et al., 1981; Tartter and Knowlton, 1981). Poizner et al. (1981) have tested which parts of the moving body were crucial for the human comprehension of ASL using PLDs with reduced sets of body markers. Their findings have suggested that the more distal the joint of the body (e.g., fingers), the more information it carries. In Tartter and Knowlton (1981), pairs of participants managed to have discussions in ASL by the only means of 27 lighting spots attached to hand articulations. These moving dots were sufficient to understand one another despite the reduced information and the increased difficulty of the task.

Additionally, some studies have further investigated the strategies of signers when perceiving visual stimuli, such as static faces or SL movements. Watanabe et al. (2011) have outlined differences in eye gaze between deaf signers and hearing non-signers during face perception. Signers focused on the eyes more frequently and longer than hearing non-signers who focused on the central face area more than the eyes. In addition to these spatial differences, Stoll et al. (2018) have shown that deaf, but also hearing, signers were slower than hearing non-signers to recognize faces, but that they recognized faces with higher accuracy. This suggests that beyond differences between deaf and hearing people, sign language acquisition influences the processing of human faces.

During the comprehension of ASL videos, eye movements have revealed that hearing beginning signers mostly fixate near signers' mouth, while deaf native signers focus on the eyes (Emmorey et al., 2009), as shown for face perception (Watanabe et al., 2011). These differences may be explained by the fact that beginners mainly relied on English mouthing (i.e., production of visual syllables with the mouth while signing). Muir and Richardson (2005) have also reported that deaf native signers fixated the facial region and used peripheral vision when viewing videos of British Sign Language (BSL). In addition to the reliance of beginning signers on mouthing for ASL comprehension (Emmorey et al., 2009), the perception of handshape and hand location compared with that of ASL signs have revealed that non-native signers predominantly focus on handshape during ASL comprehension (Morford and Carlson, 2011; Morford et al., 2008). Taken together, the studies mentioned above suggest that beginning signers concentrate on mouth movements, notably to infer information from mouthing, as well as on manual parameters, notably to process lexical information given by the hands. By contrast, native signers focus on the eyes and the facial regions while perceiving additional relevant information using peripheral vision.

2.3 The perception of identity information in SL movements: a need for further investigations

In this chapter, we reviewed visual perception and neurophysiological studies on the human perception of biological and SL motion. Using PLDs, visual perception studies have highlighted the exceptional sensitivity of human observers to the movements of their conspecifics. This sensitivity may be facilitated by the high frequency with which individuals perceive human movements in social environments, but also with the motor representations they have acquired in order to execute the same movements. Action execution and action perception may be intrisically linked,

link which could play an important role in the social-cognitive development of human beings, including the sense of self, of other, and the interaction between self and other (Lewis, 1999; Meltzoff and Moore, 1995).

Visual perception studies on PLDs have shown that this sensitivity to biological motion allows humans to identify other individuals from their movements. Human observers were able to infer identity from PLDs of a wide variety of human movements. However, although a few behavioral studies have tested the role of different classes of motion information in identification (Troje et al., 2005; Westhoff and Troje, 2007), how identity is encoded in the movements remains unclear, in particular in SL, which may reveal SL-specific patterns.

Indeed, the perception of SL motion may be distinct from that of other human movements. Not only does SL involve biological movements, but it is also constrained by linguistic rules and probably shares common processing mechanisms with spoken languages (Okada et al., 2016; Corina et al., 1999; Hickok et al., 1996). SL interestingly bridges language processing and sensorimotor integration. Likely because of this language component, the vast majority of perception studies on SL motion have investigated SL comprehension. Visual perception experiments have allowed gaining insights both into the critical motion information necessary for SL comprehension (Poizner et al., 1981; Tartter and Knowlton, 1981), and into the different perceptual strategies of native and non-native signers when processing SL discourses (Emmorey et al., 2009; Muir and Richardson, 2005).

However, to the author's knowledge, there were virtually no attempts neither to assess to which extent human observers actually manage to identify signers from their movements in SL, nor to determine the encoding of identity in the signers' movements. In addition to human perception measurements, other approaches, such as motion capture and machine learning, can provide further insights along this line.

Chapter 3

Capturing human movements

The current knowledge of the human perception of biological motion is still incomplete, notably because of its intrinsic complexity. However, recent advances in motion capture (mocap) have broken down barriers along this line, by providing tools to obtain realistic motion data in three dimensions (Section 3.1). From the raw motion data obtained with mocap systems, various higher-level features of motion can be processed (Section 3.2) whether for motion analysis, for automatic recognition of gestures or for motion generation. In particular, the advances of mocap have provided new crucial tools for research on SLs, the latter being complex to model because of the various body movements they involve (Section 3.3).

3.1 Motion capture technologies

Current mocap systems allow researchers and computer scientists to record, analyze and re-generate human movements, with high accuracy. From the beginnings of mocap to most recent advances, a wide variety of mocap systems have been developed. In the following sections, we briefly present the evolution of the mocap methods through the years (Section 3.1.1) and review the current state-of-the-art mocap technologies (Section 3.1.2).

3.1.1 Historical evolution

Motion capture (mocap) is the process of tracking the trajectories of the key points of a moving object over time. For instance, it can translate live human movements into an interpretable digitized representation (e.g., temporal vectors in three dimensions) (Menache, 2000). For more than a century, mocap has been accomplished via various techniques. In the late 1800s, scientists, including Etienne-Jules Marey, introduced the chronophotography technique (see example in Figure 3.1). Photographs capturing the successive phases of an individual's, or animal's, motion allowed them to study biological movements (Marey, 1874; Muybridge, 1887). These studies were the first investigations of locomotion, which then has been assessed with Point-Light Displays (PLDs) (see Chapter 2). A few decades after, in the film industry, mocap was achieved via rotoscoping. Tracing over an original movie, frame by frame, this technique allowed animators to produce realistic movements from drawings, such as in *Snow White and the Seven Dwarfs*, *Peter Pan* or *Alice in Wonderland*. Rotoscoping is still used, notably for the processing and generation of Sign Language (SL) motion (Segouat and Braffort, 2009). Similarly in the early 1900s, Bernstein et al. have developed efficient analyses for biomechanics based on images of light bulbs attached to the moving body, captured at high frame rates (Kay et al., 2003; Bernstein, 1927).

The widely used PLDs in motion perception studies similarly displayed moving light bulbs, which had been attached to major joints of the body (Johansson,

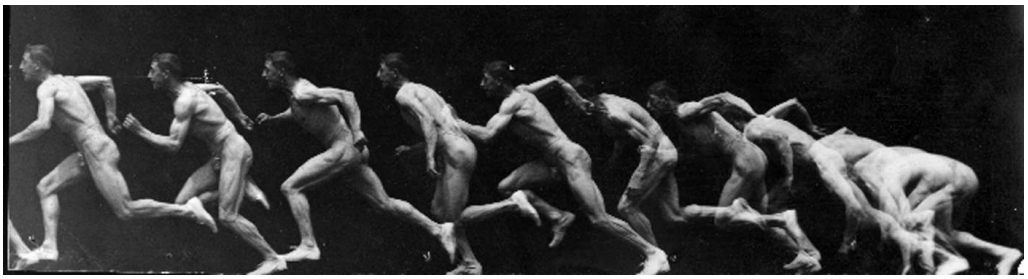


FIGURE 3.1: Example of chronophotography. A runner photographed by Georges Demenÿ (taken from Véray, 2007).

1973) (see Chapter 2 for further details). Yet, these displays rarely allowed perceiving or analyzing the movements in three dimensions, except for a few studies. Among these few studies, Poizner et al. (1981) have considered the three-dimensionality of SL production using a tachistoscope for stereoscopic presentation of videotaped PLDs. In some further computational and mathematical analyses of human movements, the trajectories of the body articulations have been manually digitized from videos (Foulds, 2004; Young and Reinkensmeyer, 2014). Recent technological advances have allowed overcoming the limitations of these frame-by-frame approaches, by providing three-dimensional recordings of human movements with high spatial and temporal accuracy.

3.1.2 Motion capture today

Since the 1980s, considerable progress has been achieved in mocap technologies. In this overview, three main categories of state-of-the-art techniques are presented: optical, mechanical and inertial systems. Further details about other techniques, such as magnetic tracking, can be found in Menache (2000).

As shown in Figure 3.2, optical systems utilize multiple cameras (e.g., 6 to 24 cameras for *Optitrack* systems) and markers placed on the moving body. The three-dimensional trajectories of the markers are obtained by triangulation using the overlapping images of the calibrated cameras. Passive optical systems (e.g., *Optitrack* or *Vicon* systems) estimate the position of the moving agent with retroreflective markers, which reflect the infrared emissions produced by the mocap system. By contrast, the markers of active optical systems (e.g., the *PhaseSpace* system) are powered to emit their own light following a specific synchronization so that the position of each marker can be identified separately. Optical systems are one of the most used mocap technologies for film and video making, but also for motion analysis in biomechanics or medicine. One main advantage of these systems is their extremely high spatial accuracy (< 1 mm). However, they often involve recovering missing data due to occlusions, light interferences or confusions between markers (Tits et al., 2018). Other camera-based techniques can be markerless, such as the Microsoft *Kinect*, which, with one camera alone, allows for the extraction of 3D trajectories thanks to a depth infrared emitter. Markerless systems have the advantage of considerably improving feasibility and ecological naturalness of mocap recordings but, up to now, they hardly provide sufficient accuracy, compared with state-of-the-art mocap systems.

Mechanical mocap systems directly track the angle data of multiple markers attached to the body thanks to a wearable body-shaped structure, called exoskeleton (see the example of *Gypsy 7* in Figure 3.3). During the movements, the system converts the analog voltage changes of the potentiometers placed on the exoskeleton's

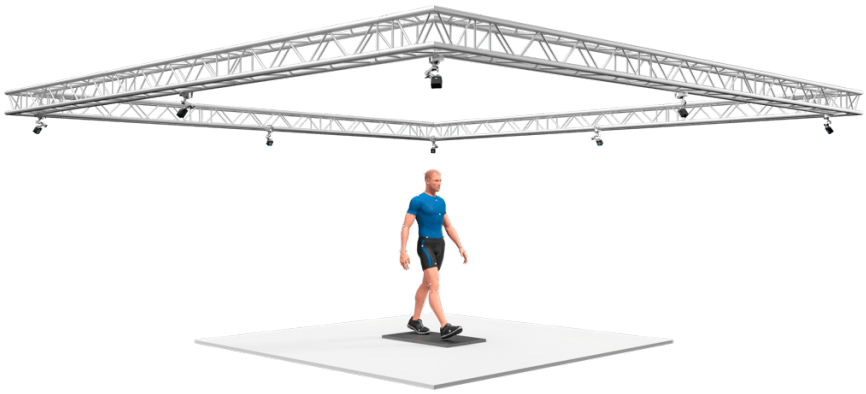


FIGURE 3.2: Example of an optical mocap system using multiple cameras and body markers (from *Optitrack*).

articulations into motion data. Although their use tends to decrease, some studies still conduct motion analyses with mechanical systems, including in SL (Malaia and Wilbur, 2012; Malaia et al., 2013). Mechanical tracking of finger movements has also been developed with gloves, such as the *Xsens gloves*, which are provided with sensors detecting the flex of finger joints. Gloves have been used extensively for automatic SL recognition (Grimes, 1983; Fels and Hinton, 1993; Liang and Ouhyoung, 1998; Oz and Leu, 2011; Saggio et al., 2020). However in SL, these systems, often built by hearing teams, do not reflect real-world use cases. For instance, gloves only focus on hands while SL is a continuous stream of various motion features, including hand gestures, but also torso movements, facial expressions or eye gaze (Erard, 2017).

FIGURE 3.3: Example of the *Gypsy 7* mechanical mocap system.

Like mechanical systems, inertial mocap systems do not involve cameras. These systems rely on Inertial Measurement Units (IMUs), which combine miniature sensors, including accelerometers, gyroscopes and magnetometers (Xsens, 2021). As shown in Figure 3.4, IMUs are wireless and can be attached to key body joints via wearable suits, as for optical mocap technologies. The IMU sensors provide raw measures of linear acceleration, angular velocity and global orientation, which are interpreted by the software and mapped to a skeleton using biomechanical models

and sensor fusion algorithms. Because of their portability and ability to communicate wirelessly, inertial and accelerometer-based systems are often used for the development of motion interfaces, notably for the analysis and control of musical gestures (Schoonderwaldt et al., 2006; Bevilacqua et al., 2007; Rasamimanana et al., 2010).

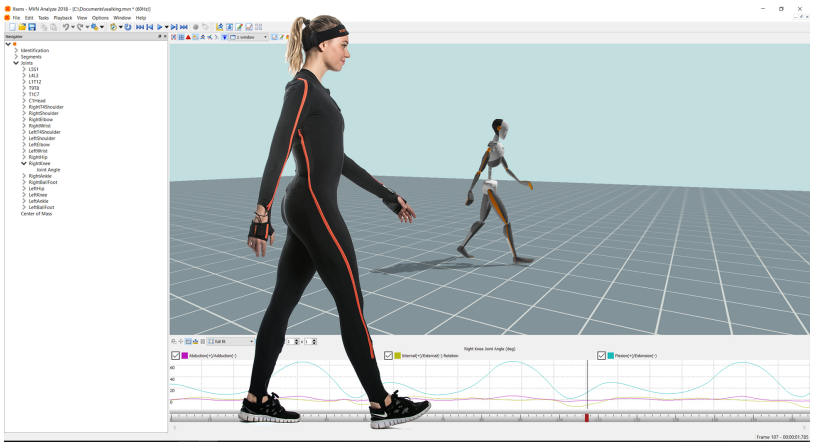
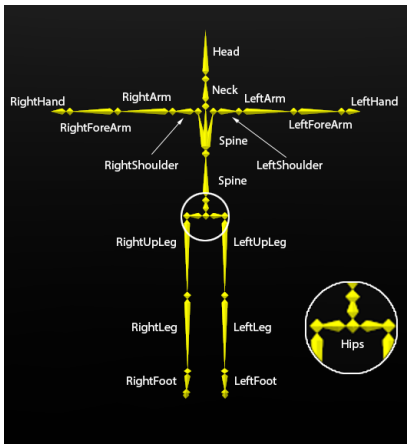


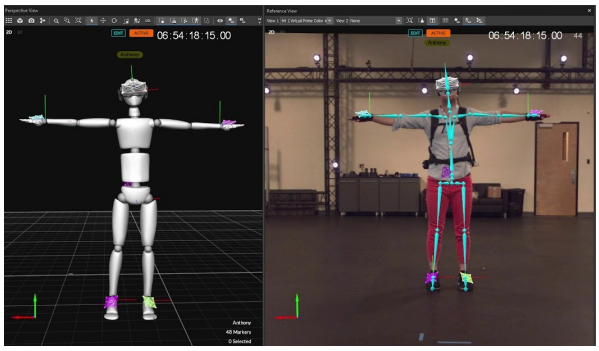
FIGURE 3.4: The *Xsens Mtw Awinda* inertial mocap equipment and its software *MVN Animate*.

Among the mocap techniques mentioned above, researchers must assess the trade-off between accuracy and portability. For instance, accelerometer-based systems allow recording movements in various environments and in larger areas than optical systems. However, the state-of-the-art precision of optical systems and their ability to record full-body motion make them better suited for accurate analyses of human movements. For these reasons, the mocap data used in the present thesis were obtained using an *Optitrack* mocap system, equipped with optical passive markers and 10 cameras with a spatial resolution under 1 mm and a temporal resolution of 250 fps.

3.2 Motion representations: from body markers to high-level features



(A)



(B)

FIGURE 3.5: Motive software (*Optitrack*): (A) Body joints structure of the skeleton proposed, (B) mapping of the markers positions and orientations to the skeleton. Images were taken from the *Optitrack* documentation and *Motive 2.1 | What's New*.

The mocap technologies reviewed above can provide measures of positions (Section 3.2.1) and angles (Section 3.2.2) in different manners. Positions and angles are measured at key articulations of the body. They can be processed either independently for each marker or following a hierarchical structure defined by a standardized skeleton, often provided by mocap softwares (e.g., Motive for *Optitrack*, as shown in Figures 3.5a and 3.5b). Further descriptors of the movements, such as kinematic and kinetic features (Section 3.2.3) or Principal Movements (Section 3.2.4), can then be derived from these raw data for specific motion analyses. The latter higher-level motion descriptors have been the basis of most studies investigating the extraction of human characteristics from biological movements (Troje, 2002a), including identity (Zhang and Troje, 2005; Carlson et al., 2020).

3.2.1 Position data and the impact of anthropometrics

Measures of position of the body markers are given by the 3D Cartesian coordinates estimated by triangulating the images of the different cameras involved in the mocap system. This representation allows for the visualization of human movements using stick figures, or PLDs. Indeed, most recent studies investigating the human perception of biological motion have constructed their PLDs using mocap recordings (Troje, 2002a; Westhoff and Troje, 2007; Bläsing and Sauzet, 2018; Sevdalis and Keller, 2009; Baragchizadeh et al., 2020). In further computational analyses of motion, these position data are often converted into a reference system centered on a root marker, such as the pelvis (Carlson et al., 2020), the body Center of Mass (CoM) (Zago et al., 2017a) or more specific key points related to the task (Federolf et al., 2014) (see an example in Figure 3.6). In most cases, these derived references attached to the moving persons are more appropriate than stationary external references, notably for the analysis of multiple motion examples performed by different persons who may have been oriented differently in the capture area.

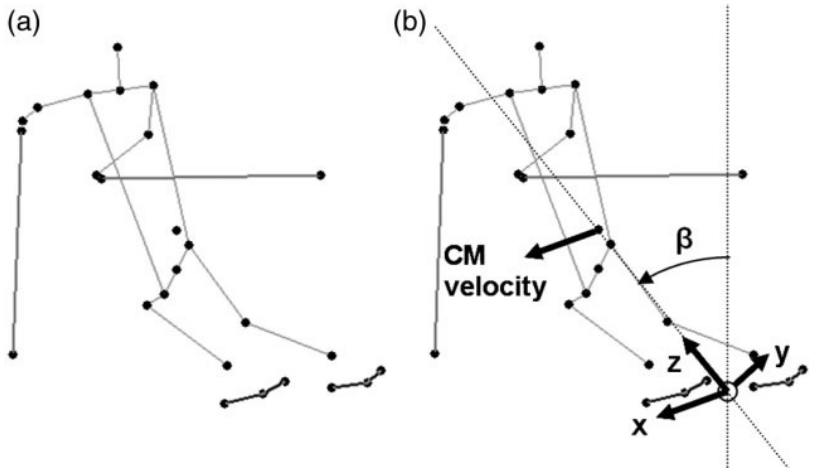


FIGURE 3.6: Example of body-centered reference system, with the midpoint between the two skis of the subject as origin (Federolf et al., 2014).

Either with external or body-centered references, 3D Cartesian coordinates allow reconstructing a moving skeleton, but they include anthropometric measures (e.g., height or shoulder width) specific to each individual. Some methods have been proposed to filter out anthropometric differences in mocap datasets of multiple persons. The coordinates of body markers can be scaled by dividing them by a reference length, such as the torso height or the distance between the head and the

origin (e.g., pelvis) (Sie et al., 2014; Morel et al., 2016). In other studies, the comparison between the motion of different individuals was eased by a normalization of the average mocap postures. This was achieved by subtracting the average posture of each individual from their mocap postures (i.e., at each frame), and replacing it with the mean norm of average postures across all individuals (Troje et al., 2005; Westhoff and Troje, 2007; Zago et al., 2017a; Federolf et al., 2013b). Additionally, Troje et al. (2005) have introduced a method to make the mocap data of multiple walkers share the same body height while keeping information about the shape of their body (e.g., shoulder width) intact, using linear regression. In Chapter 5 of this thesis, we replicate these techniques, propose further normalization steps and apply them to LSF mocap data.

The methods mentioned above have proposed simple mathematical transformations of mocap data in order to decrease the effect of anthropometric differences in motion analyses. Other morphology-related influences can correlate with differences in the movements of multiple individuals. For instance, the weight and height of one's body may influence the kinematics of his or her movements. Taking this effect into account, Tits et al. (2017) have proposed a method to directly remove the influence of morphology-related factors on motion features (e.g., speed, acceleration peaks) via linear regression. By contrast with scaling approaches, this latter method can be generalized to any morphology-related factor and, more importantly, directly manipulates the effect of morphology on the high-level features.

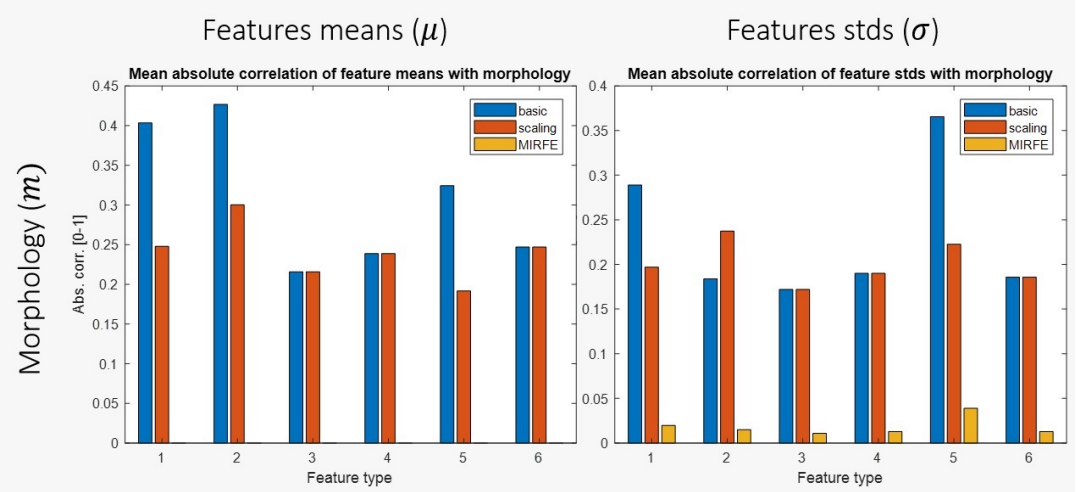


FIGURE 3.7: Results reported by Tits et al. (2017) using their method (MIRFE) for morphology-independent processing of mocap data. Their method drastically reduces the correlation between motion features statistics and morphology-related factors, compared with classical scaling methods, which reduce it only partly.

Considering the 3D Cartesian coordinates of each body marker independently thus increases the impact of anthropometric factors. Moreover, the lack of constraints for the segment lengths can be problematic for some applications, such as motion synthesis. Unless imposing the size invariance of the segments in the synthesis algorithm, the generated animations can produce movements with segments varying their sizes, which is likely to be unrealistic. To overcome these two limitations, a standardized skeleton can be used with fixed segment lengths, irrespective of the anthropometric measures of the moving person. In most cases, these skeleton mappings require further information in addition to position data, notably about the orientations and angles of the body joints.

3.2.2 Angular data

In most cases, the standardized skeleton used when capturing human motion is a hierarchical chain of body articulations with fixed segment lengths whose movements are described as 3D rotations with respect to their preceding articulation (recursively toward the root marker) (see Figure 3.5a). A 3D rotation can be described using multiple rotation representations, such as Euler representation. The three Euler angles represent three successive rotations with respect to the axes of the 3D coordinate system $Oxyz$. For instance, in the rotation given by the Euler angles (ϕ, θ, ψ) (Figure 3.8), ϕ corresponds to the angle of the first rotation around the z-axis. θ corresponds to the angle of the second rotation around the x'-axis (former x-axis). ψ corresponds to the angle of the third rotation around the z'-axis (former z-axis). The new rotated coordinate system is $Ox'y'z'$.

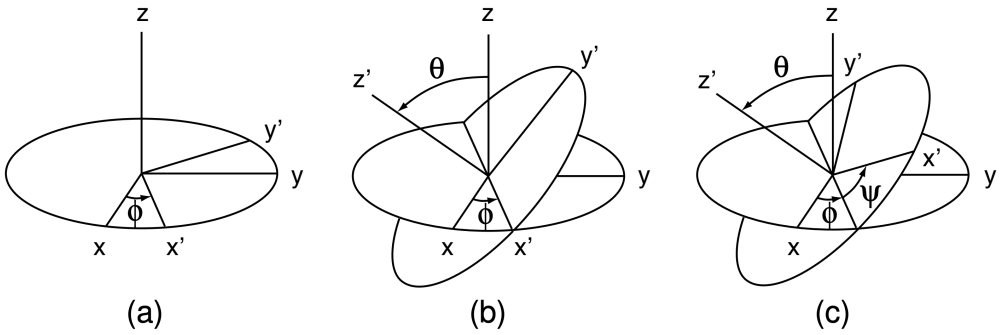


FIGURE 3.8: The successive steps of rotation for Euler angles (Taken from Schwab and Meijaard (2006)).

The 3D rotation can also be described using a 3×3 rotation matrix. For instance, the full rotation shown in Figure 3.8 can be seen as the matrix product of three rotation matrices, corresponding to the successive rotations described above:

$$\mathbf{A} = \mathbf{BCD} \tag{3.1}$$

where $\mathbf{A}$ corresponds to the resulting 3D rotation and $\mathbf{D}, \mathbf{C}, \mathbf{B}$ corresponds to the successive rotations around the z-, x'- and z'-axes, respectively.

The parameters of the rotation matrix $\mathbf{A}$ can be found using Euler angles, as each of the matrices $\mathbf{D}, \mathbf{C}$ and $\mathbf{B}$ are described by their respective Euler angles:

$$\mathbf{D} = \begin{bmatrix} cos(\phi) & sin(\phi) & 0 \\ -sin(\phi) & cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta) & sin(\theta) \\ 0 & -sin(\theta) & cos(\theta) \end{bmatrix} \tag{3.3}$$

$$\mathbf{B} = \begin{bmatrix} cos(\psi) & sin(\psi) & 0 \\ -sin(\psi) & cos(\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.4}$$

Therefore, although the 3D rotation matrix **A** is defined by nine parameters, it can easily be described using three parameters: the Euler angles. The low dimensionality of the Euler representation allows for successful fast calculation of rotational motion. However, the major drawback of the Euler representation is its ambiguity. For instance, the three Euler angles can be applied to different axes, depending on the convention used (we used the "x-convention" above). Moreover, even using a fixed convention, different combinations of Euler angles can refer to the same rotation. This rotation representation is thus not unique. For these reasons, the quaternion representation (Hamilton, 1866) can be preferred to Euler angles. Quaternions are similarly low-dimensional, as they are defined by four parameters (i.e., one scalar and three complex components), and they do not suffer from ambiguity, as they represent a unique rotation.

An extensive description of the quaternion representation is out of the scope of the present thesis, in particular as most of the prior work that formed the theoretical foundations of this thesis do not rely on this representation. Indeed, quaternions, and rotation representations in general, are mostly used for motion synthesis and animation applications, in order to produce realistic movements (Tilmanne and Dutoit, 2010; Tilmanne et al., 2014; Tilmanne et al., 2012; Felis et al., 2015; Brand and Hertzmann, 2000; Alemi et al., 2015). By contrast, motion analysis studies mostly rely on 3D Cartesian coordinates, and more specifically on higher-level (e.g., kinematic and kinetic) features processed from these coordinates.

3.2.3 Kinematic and kinetic features

A wide majority of motion analyses have assessed the kinematic properties of the movement. Kinematics refer to the aspects of motion, without considering masses or forces involved in it. For instance, the instantaneous linear velocity and acceleration of body markers are often derived from their positions:

$$\mathbf{p}(t) = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \tag{3.5}$$

where $\mathbf{p}(t)$ is the 3D position vector of a given body marker and x, y, z are the 3D Cartesian coordinates along the **i**, **j**, **k** unit vectors of the 3D reference system.

$$\mathbf{v}(t) = \frac{d\mathbf{p}}{dt} = \frac{dx}{dt}\mathbf{i} + \frac{dy}{dt}\mathbf{j} + \frac{dz}{dt}\mathbf{k} \tag{3.6}$$

where $\mathbf{v}(t)$ is the linear velocity vector.

$$\mathbf{a}(t) = \frac{d\mathbf{v}}{dt} = \frac{d^2x}{dt^2}\mathbf{i} + \frac{d^2y}{dt^2}\mathbf{j} + \frac{d^2z}{dt^2}\mathbf{k} \tag{3.7}$$

where $\mathbf{a}(t)$ is the linear acceleration vector.

These linear kinematic features are frequently used in the motion analysis of SLs (Catteau et al., 2016; Blondel et al., 2019; Malaia and Wilbur, 2012; Malaia et al., 2013). Analogous computations can be done to quantify the kinematics of rotating objects, by deriving the angular velocity and acceleration of the object from its angular positions. Higher-order derivatives of position can also be of interest. For instance, jerk

quantifies the rate at which an object's acceleration changes with respect to time. It has been used to assess the smoothness of movements, notably in musical conducting (Sarasúa and Guaus, 2014). Moreover, kinematic features can be interpreted using their average, rather than instantaneous, values. For instance, the average velocity of all body markers can be computed to account for the overall Quantity of Motion (QoM), often used for the analysis of dance or musical movements (Sarasúa and Guaus, 2014; Camurri et al., 2003b; Camurri et al., 2003a).

In addition to kinematic features, kinetic measures can also be used to assess movements (Toiviainen et al., 2010; Reid, 2010; Winter, 2009; Luck et al., 2009). Unlike kinematics, kinetics assess the relationships between the movement and its causes, such as masses and forces applied to it. For instance, the kinetic energy of a body is the energy due to its motion. It depends on the speed and mass of the body. First, the position of the body Center of Mass (CoM) (i.e., the mass-weighted average of the positions of body segments) is calculated (equation 3.8) (Winter, 2009). Second, the kinetic energy is computed based on the overall mass of the body and on its translational and rotational velocity calculated at the CoM location (equation 3.9) (Winter, 2009):

$$\mathbf{p}_{CoM} = \frac{1}{M}\sum_{i=1}^{N} m_i \mathbf{p}_{CoM_i} \tag{3.8}$$

where $\mathbf{p}_{CoM}$ is the position vector of the body CoM, M is the total mass of the body, m_i and $\mathbf{p}_{CoM_i}$ are the mass and CoM position vector of body segment i, respectively, N is the total number of body segments.

$$E_{kin}(t) = E_{trans}(t) + E_{rot}(t) = \frac{1}{2}Mv_{CoM}^2(t) + \frac{1}{2}I\omega^2(t) \tag{3.9}$$

where E_{trans} and E_{rot} are the translational and rotational kinetic energy, respectively, M is the total mass of the body, v_{CoM} is the linear velocity of the body CoM, I is the moment of inertia of the body and ω is the angular velocity of the body CoM.

3.2.4 Principal Movements

As mentioned above, from the raw mocap data, motion studies can use specific higher-level (e.g., kinematic or kinetic) variables of the movements, defined by researchers. Some studies have taken another approach to describe movements from a holistic perspective, using Principal Component Analysis (PCA). Unlike the analysis of pre-selected variables, this data-driven method has allowed disentangling how complex multi-segmental movements were structured, without any a priori hypotheses. Troje (2002a) first used PCA to extract motion patterns allowing for the gender classification of gait. Similarly to "eigenfaces" (O'Toole et al., 1993) or "egeinvoices" (Kuhn et al., 2000), the whole walker's movement was decomposed into simpler one-directional principal movements (PMs) (i.e., time series of "eigenpostures"), which maximized the variance in the original motion (for further details about PM decomposition, see the Methods section in Chapter 7). Based on the temporal characteristics of a reduced set of PMs, a linear classifier was then able to predict the gender of the walker. Similar studies have also reported that PMs allowed for automatic prediction of a walker's identity (Zhang and Troje, 2005) and mental state (Sigal et al., 2010).

Concerning other movements, Federolf et al. (2014) have introduced this decomposition method for the evaluation of an athlete's technique, taking skiing as an example. Using the invertibility of PCA, PMs were projected back onto the original

3D space and were visualized (Figure 3.9). This allowed interpreting and comparing the skiing movements of athletes in terms of distinct PMs, such as lateral body inclination, flexion-extension of the legs or rotation of the skis.

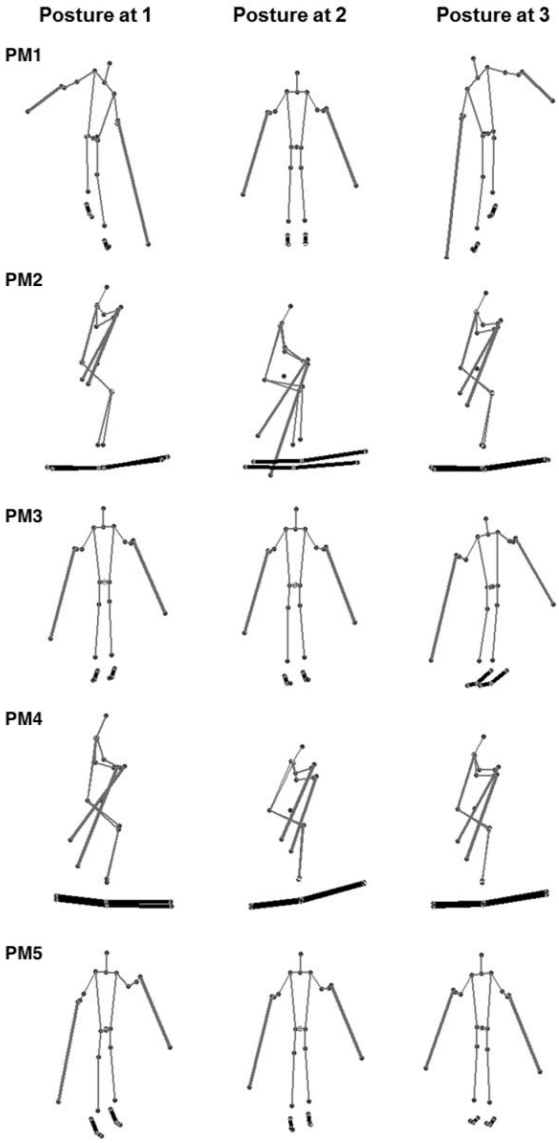


FIGURE 3.9: The first five Principal Movements (PMs) in skiing, reported by Federolf et al. (2014)). Postures at 1, 2 and 3 represent the PM at the time instants corresponding to a large positive, small, or large negative PM weighting, respectively.

The same method has been successfully applied to other sports. Combined with other features, PMs have allowed for the automatic estimation of karatekas' experience (Zago et al., 2017a), or for the automatic evaluation of dives (Young and Reinkensmeyer, 2014). In juggling, individual differences due to experience have been found in specific PMs (Zago et al., 2017b). Moreover, such differences in PMs have allowed determining discriminant patterns between healthy and pathological gait (Federolf et al., 2013a; Zago et al., 2017c). Several studies have also used this technique to investigate human posture control (Federolf et al., 2013b; Federolf, 2016; Longo et al., 2019; Haid et al., 2018; Berret et al., 2009). More specifically, Haid et al. (2018) have reported age effects in postural control characterized by control

differences in specific PMs. In the artistic domain, Tits et al. (2015) have showed that the finger gestures of pianists can be decomposed into eight PMs and that the complexity of the decomposition was a function of pianists' expertise.

3.3 Motion capture corpora in Sign Language

Motion is a crucial component of SLs, as they rely on a visual-gestural modality. SL corpora thus combine information of various types, such as videos, pictures, annotations or 3D motion data, in order to describe the movements involved in SL productions. They can be used for various purposes, such as motion analysis, linguistics or computer science. Beyond the first corpora made using video recordings (Section 3.3.1), the advances of motion capture have allowed building SL corpora with 3D motion data (Section 3.3.2). These corpora have allowed providing tools for SL automatic processing (Section 3.3.3), as well as for the fundamental analysis of SL motion.

3.3.1 Video corpora

Up to now, most of SL corpora have been constructed using video recordings. Video recording, whether using one or more cameras, is the simplest and most low-cost form of production of SL corpora. Videos have allowed for the production of SL lexicons, such as the American Sign Language Lexicon Video Dataset (ASLLVD) (Neidle et al., 2012). In this corpus, 2284 lexical American Sign Language (ASL) signs were recorded, using multiple cameras at different resolutions and frame rates (i.e., 1600 × 200 at 30 fps and 640 × 480 at 60 fps), and from different viewpoints (i.e., frontal and profile views) (Figure 3.10). In MS-ASL, Joze and Koller (2018) have gathered up to 1000 ASL signs using Youtube videos of ASL lessons. Video lexicons have been proposed for other SLs. For instance, the Signum corpus (Von Agris and Kraiss, 2007) provides 450 lexical signs of German Sign Language (DGS) recorded using RGB cameras, with a 776 × 578 resolution at 30 fps.

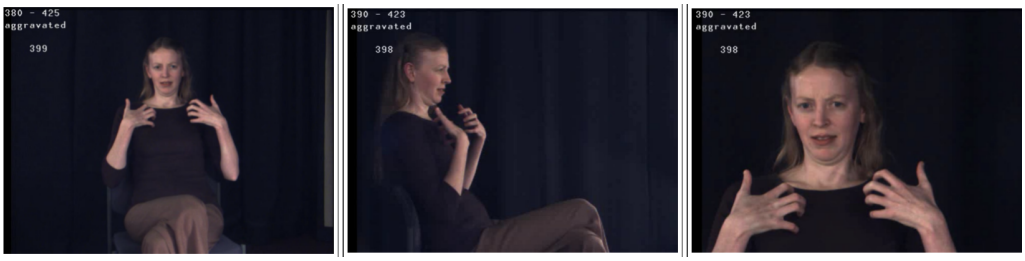
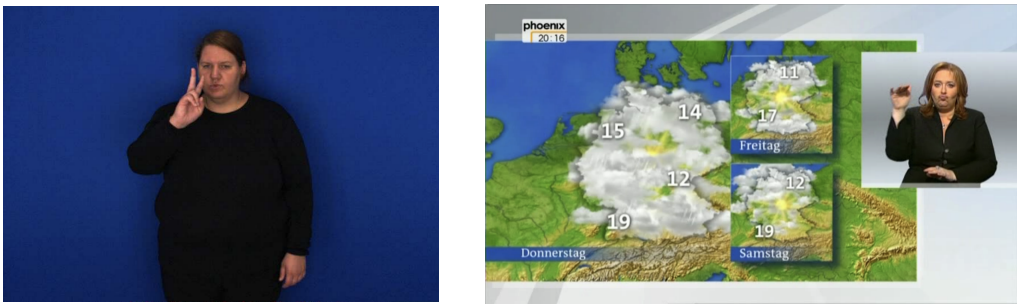


FIGURE 3.10: The different viewpoints from which the ASLLVD corpus was recorded (Neidle et al., 2012).

In addition to isolated lexical signs, the DGS Signum corpus also incorporates continuous SL (e.g., DGS sentences formed with the isolated signs mentioned above) (Von Agris and Kraiss, 2007) (Figure 3.11a). Unlike spoken languages, which can be seen as concatenations of words, spontaneous SLs are continuous and complex streams of multiple movement patterns. Corpora of isolated signs thus only partly account for the production of SLs in real life. For this reason, some SL corpora have aimed to include continuous SL in their recordings. For instance, in the RWTH Phoenix corpus (Forster et al., 2014; Koller et al., 2017), the spontaneous productions of weather forecasting in DGS were recorded for multiple interpreters, using

an RGB camera at 25 fps and with low spatial resolution (210 × 260 pixels) (Figure 3.11b). The Auslan corpus (Johnston et al., 2009) comprises 300 hours of naturalistic Australian Sign Language (Auslan), recorded on digital videotapes. Video corpora of continuous SL are also available in ASL. For instance, the Purdue RVL-SLL corpus (Martínez et al., 2002) contains ASL productions of multiple signers recorded on videotapes, including isolated lexical signs as well as short discourse examples.



(A) (B)

FIGURE 3.11: Examples of the video recordings provided by SL video corpora. (A) Signum corpus in DGS (Von Agris and Kraiss, 2007), (B) RWTH Phoenix corpus in DGS (Forster et al., 2014).

Other continuous SL corpora have been released in different SLs (e.g., BSL corpus in British Sign Language (Vinson et al., 2008), LSFB corpus in French Belgian Sign Language (Meurant and Sinte, 2016), NGT corpus in Dutch Sign Language (Crasborn and Zwitserlood, 2008)). In some cases, multiples SLs are also incorporated in the same corpus. For instance, the Dicta-Sign corpus (Efthimiou et al., 2010) have gathered videos of SL productions in four different SLs: British (BSL), Greek (GSL), German (DGS) and French (LSF). For a more complete list of SL video corpora, see surveys by Reiner Konrad[1] and Koller (2020).

Moreover, some SL video corpora, such as the RWTH Boston corpus (Zahedi et al., 2006), have used multiple cameras, in particular pairs of cameras in frontal view in order to provide stereo recording of the signers, which can allow obtaining three-dimensional data. Indeed, one major limitation of video corpora is that no 3D information about the movements is provided when using RGB videos. As shown in Figure 3.10 for ASLLVD (Neidle et al., 2012) and in Figure 3.12 for RWTH Boston (Zahedi et al., 2006), some video corpora include recordings in both frontal and profile views, which allow gaining insights into the SL movements along mediolateral, anteroposterior and vertical directions. However, no accurate 3D information is available from video corpora, which can be a consequent limitation for SL applications. For instance, the automatic recognition of SL can be quite challenging without any information about depth (e.g., for distinguishing between different handshapes in 3D). To overcome this limitation, an increasing number of SL corpora have taken advantage of the advances of motion capture in order to record 3D data of SL movements.

3.3.2 From videos to 3D full-body motion capture

Some SL corpora have used depth cameras, such as the Microsoft *Kinect*, which allow reconstructing the 3D positioning of the main body articulations. For instance,

[1] https://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/SL-Corpora-Survey_update_2012.pdf

FIGURE 3.12: The two different viewpoints (top: frontal / bottom: profile) from which the RWTH Boston corpus was recorded, for three different signers (Zahedi et al., 2005).

the DEVISIGN (Chai et al., 2014) and CSLR (Huang et al., 2018) corpora in Chinese Sign Language have used Microsoft *Kinect* cameras in order to record isolated lexical signs and continuous SL, respectively. Using these cameras, both corpora include a combination of RGB and depth images. Depth cameras have also allowed for the recording of SL corpora in Greek and German Sign Language (Cooper et al., 2012) (Figure 3.13) as well as in Polish Sign Language (Oszust and Wysocki, 2013). The 3D trajectories of the body articulations can be reconstructed from the depth data, which has allowed improving the performance of SL recognition models (Pu et al., 2016). Dilsizian et al. (2016) further have shown the importance of 3D trajectories for distinguishing between signs of similar handshapes, by training an SVM (Support Vector Machine) classifier that successfully recognized ASL signs from Microsoft *Kinect* data.

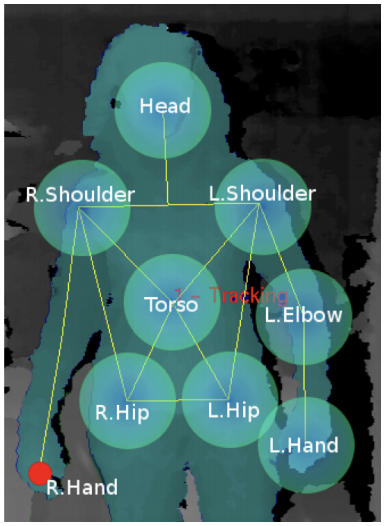
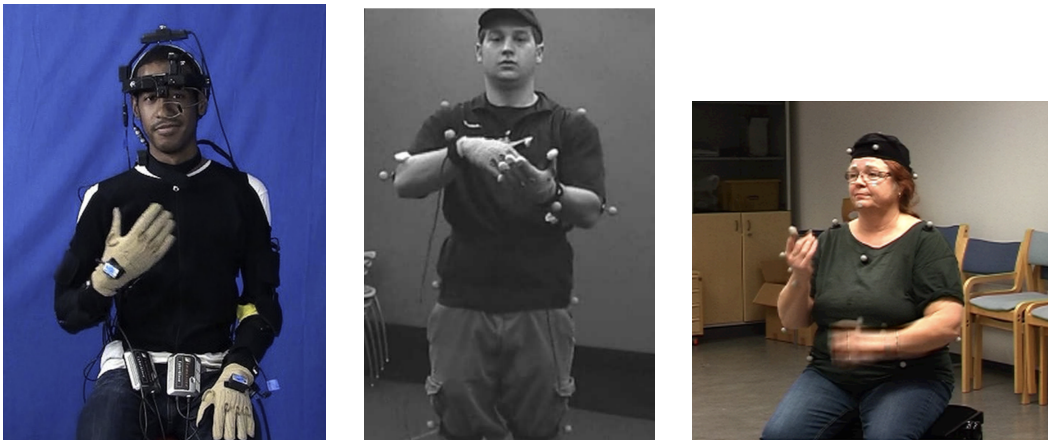


FIGURE 3.13: Body articulations used to extract the 3D motion trajectories of signers, from Microsoft *Kinect* recordings, in Cooper et al. (2012).

It thus appears essential that computational models of automatic SL processing take 3D information into account. For this reason, state-of-the-art full-body mocap

systems have been of particular interest in order to develop SL corpora with 3D motion data. Their limited use in SL corpora so far, compared to that of simple video cameras, can be explained by practical factors. Unlike RGB and depth setups, these systems involve a high number of cameras as well as numerous markers to be fixed on the signer. Moreover, mocap systems, such as those of *Optitrack* or *Vicon*, are expensive and require specific logistical conditions in order to achieve accurate 3D recordings (e.g., enough room for the multiple infrared cameras, controlled lighting to avoid interferences). Still, state-of-the-art 3D mocap systems allow obtaining highly accurate recordings of the 3D movements of signers (e.g., with a spatial resolution under 1 mm and at high sampling rates, such as 100 or 250 fps) and thus have opened up promising perspectives toward the collection of SL corpora with 3D motion data.

In ASL, Lu and Huenerfauth (2010) have used multiple 3D mocap systems to record natural discourses produced by signers for the CUNY corpus (Figure 3.14a). The inertial mocap systems *Animazoo IGS-190* and *Intersense IS-900* were used to obtain the 3D trajectories of the upper-body joints (i.e., wrists, elbows, shoulders, clavicle, neck and waist) and of the head of the signer, respectively. Additionally, two sensor gloves recorded finger movements and one eye tracker recorded the signer's eye gaze direction. In its latest version, the CUNY corpus comprised ASL productions of eight signers (Lu and Huenerfauth, 2014). For more specific analyses, Malaia et al. (2008) have recorded the 3D movements of one signer producing over 50 verb signs in ASL, using a *Gypsy* mechanical mocap system (Figure 3.14b). In Tyrone et al. (2010), the 3D movements of the head, torso and arms of multiple signers producing ASL discourses were recorded by means of a *Vicon* optical mocap system.



(A) (B) (C)

FIGURE 3.14: Examples of the 3D mocap systems used for the collection of SL corpora. (A) CUNY 3D ASL corpus (Lu and Huenerfauth, 2014), (B) 3D mocap data of ASL verb signs (Malaia et al., 2008), (C) 3D mocap corpus in Finnish Sign Language (Jantunen et al., 2012).

Optical mocap systems have also been used to collect 3D motion data in Finnish Sign Language (Jantunen et al., 2012) (Figure 3.14c) and French Sign Language (LSF) (Duarte and Gibet, 2010; Heloir et al., 2005; Catteau et al., 2016; Benchiheub et al., 2016b). More specifically in Heloir et al. (2005), a 3D motion acquisition protocol has been proposed to record LSF movements using a combination of an optical *Vicon* mocap system and two sensor gloves. In Catteau et al. (2016), a *Vicon* mocap system allowed recording the 3D movements of one signer when producing poetic

LSF sequences. In the present thesis, we used the MOCAP1 corpus, an LSF corpora in which the 3D movements of eight signers were recorded using an optical *Optitrack* mocap system (Benchiheub et al., 2016b). The movements of various body parts (including movements of the head, arms, hands and upper-body as well as facial expressions) of the eight signers were recorded during the description of different pictures in spontaneous LSF. For further details about the MOCAP1 corpus, see Chapter 5.

3.3.3 What for?

One major application of motion capture in SL is automatic SL processing, that is automatic SL recognition, translation and generation. The first studies in automatic SL recognition have relied on video recordings and image processing methods, notably for the recognition of isolated signs (Tamura and Kawasaki, 1988; Grobel and Assan, 1997; Charayaphan and Marble, 1992). For instance, Tamura and Kawasaki (1988) have built a computational system capable of classifying Japanese Sign Language isolated signs from motion features computed from the signer's hands regions. To do so, the hands regions were extracted using skin color thresholding on the videotapes. Similar hand tracking approaches, whether using skin color detection methods or colored gloves, have allowed for quite accurate video-based recognition of continuous ASL (Starner and Pentland, 1997). In addition to these studies focusing on handshapes, some accurate video-based SL recognition systems have been proposed using further bodily motion features, such as full upper-body pose or facial expressions (Forster et al., 2014).

As mentioned in Section 3.3.2, 3D trajectories can be crucial for the automatic recognition of SL (e.g., for recognizing signs of similar handshapes (Dilsizian et al., 2016)). Focusing on the important role of finger movements in SLs, in particular for lexical signs or fingerspelling (i.e., spelling out isolated words using the manual representations of letters of the alphabet, mainly used for proper nouns with no signed equivalent), most studies first used sensor gloves to consider three-dimensional motion data of the signers' fingers (Grimes, 1983; Fels and Hinton, 1993; Liang and Ouhyoung, 1998; Oz and Leu, 2011). For instance, the Glove-Talk system (Fels and Hinton, 1993) accurately produced spoken words from the automatic recognition of 203 hand gestures (less than 1% error rate), many of them being derived from the ASL. Gloves have also been used to automatically recognize LSF signs from lexicons as well as from continuous utterances (Braffort, 1996). Similarly in Liang and Ouhyoung (1998), isolated signs and sentences in Taiwanese Sign Language were automatically recognized in real-time, using a sensor glove on the dominant hand of the signer, with an average accuracy of 80.4%. In the latter study, a 3D tracker of hand orientation was used jointly with the sensor glove, similarly to Oz and Leu (2011) in ASL. Note that most studies mentioned above concern ASL, which particularly involves hand and finger movements, such as in fingerspelling. Moreover, even in ASL, neither can SL be restricted to hand gestures nor to fingerspelling. In addition to hand gestures and fingerspelling, SL production involves multiple features, including rapid manual (e.g., pointing) and non-manual (e.g., eye gaze) movements.

In that regard, depth cameras, as well as being non-intrusive compared to 3D body sensors, have allowed developing SL recognition systems using further body features. Even though depth cameras still have been used in some studies to focus on hand gestures (Lang et al., 2012; Uebersax et al., 2011), they also have allowed recognizing SL discourses from 3D motion features of other body joints, such as elbows, shoulders and neck (Zafrulla et al., 2011). SL recognition systems have also

been proposed using a combination of multiple cameras and accelerometers, in order to track multiple 3D motion features, such as movements of the wrists and torso (Brashear et al., 2003). However, the recent impressive advances of machine learning, in particular deep learning and Convulational Neural Networks, have allowed obtaining convincing recognition systems based on video corpora, rather than using 3D data from glove sensors, depth cameras or accelerometers (Cui et al., 2019; Koller et al., 2018; Shi et al., 2018; Koller et al., 2019; Forster et al., 2014; Belissen et al., 2020). For further details about automatic SL recognition, including extensive descriptions of motion extraction techniques and of recognition methods, see Koller (2020).

The second main field of automatic SL processing is automatic SL translation. For instance, some systems use automatic SL recognition and then translate its output into written representations. Up to now, neither do SLs have a written form nor a graphic system. Still, notation systems, such as SignWriting (Sutton, 2009) or Hamnosys (Hanke, 2004), have been proposed to represent SLs. Some other studies have used deep learning methods to develop translation systems that automatically recognize SL discourses from videos and provide their translations into spoken languages (i.e., sign-to-speech) (Koller et al., 2016; Koller et al., 2018; Camgoz et al., 2018). In addition to the translation of SL into written and spoken forms, an important part of research in SL processing has focused on the translation of written and spoken languages into SL (i.e., text-to-sign and speech-to-sign) (Davydov and Lozynska, 2017; Elliott et al., 2000; Veale et al., 1998; Zhao et al., 2000; Karpouzis et al., 2007; Ebling and Glauert, 2016). As shown in Figure 3.15, translation machines from spoken languages to SL thus involve the generation of SL animations using virtual signers (or signing avatars).

(A) (B)

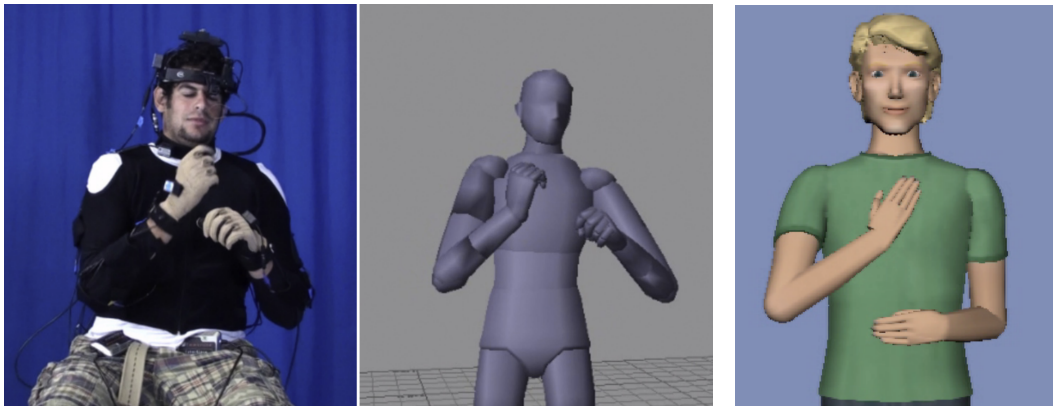
FIGURE 3.15: ViSiCAST project for the automatic translation from text to SL (Elliott et al., 2000). (A) Signer wearing the 3D mocap setup during the recording of SL discourses or isolated signs, (B) Example of virtual signer to which the mocap SL data can be mapped.

Automatic SL generation and virtual signers allow producing accessible content in SL for deaf individuals whose first language is an SL and whose comprehension of written content is not always mastered. Moreover, the use of virtual signers brings many additional advantages, by contrast with pre-recorded videos of signers: it is possible to reuse, adapt, or modify the content of the animation; the appearance of avatars (age, gender, style of clothing, etc.) can be modified according to the target population; they can be dynamic and interactive (Kipp et al., 2011). Virtual signers are thus particularly suited for automatically generating dynamic content (e.g., for journalistic websites or TV announcements). Compared to automatic recognition, whose state-of-the-art models mostly rely on video datasets, an increasing number of SL studies use 3D motion capture data to provide realistic and comprehensible virtual signers animations. Indeed, although 3D mocap is not the only method used to animate virtual signers (e.g., key-frame animations for the Paula system shown in Figure 3.16 (Filhol et al., 2017; Filhol and Mcdonald, 2020)), most 3D mocap SL corpora presented in Section 3.3.2 were built for animation purposes (Lu and Huenerfauth, 2010; Lu and Huenerfauth, 2014; Duarte and Gibet, 2010; Gibet, 2018; Naert et al., 2020) (Figure 3.17).

In summary, capturing SL motion using either 2D videos or 3D mocap recordings has allowed developing tools for the automatic processing of SL. While deep learning techniques have improved state-of-the-art SL automatic recognition systems from 2D videos, 3D mocap data have a crucial role for the animation of realistic and comprehensible virtual signers, as they allow replaying real human movements on a controlled virtual agent. However, tools for automatic SL processing are far from being as effective and as deployed as for spoken languages, notably because of the lack of SL datasets but also because of the intrinsic complexity of the movement features involved in SLs. There is still much to learn about the characteristics of SL movements in order to produce convincing synthetic animations. For this purpose, the 3D motion capture of SL movements brings another substantial contribution to SL research: motion analysis. The advances of 3D mocap techniques jointly with those of computational methods, including machine learning, have allowed analyzing complex movements with high accuracy and thus gaining insights into the characteristics of SL movements for various purposes (Benchiheub et al., 2016b), such as linguistics, motion science, visual perception as well as for technological applications, such as motion generation for virtual signers.

(A) (B)

FIGURE 3.16: Animation of the Paula system for two different SL expressions (Filhol and Mcdonald, 2020).



(A) (B)

FIGURE 3.17: The CUNY ASL Corpus for the animation of virtual signers (Lu and Huenerfauth, 2010; Lu and Huenerfauth, 2014). (A) Signer wearing the 3D mocap setup during the recording of ASL discourses (left) and the corresponding avatar skeleton (right), (B) The resulting virtual signer animated using the 3D mocap data.

Chapter 4

Motion analysis and machine learning

Machine learning models are computational models able to learn from amounts of data in order to predict some output variable from unknown data. By optimizing a mathematical model, machine learning models are able to extract specific aspects of the input data and to link them with the desired output (i.e., learning). This mathematical relationship between inputs and outputs is then used by the model in order to provide accurate predictions from novel data. Machine learning models have been widely used from audio data for various automatic applications, such as speech recognition (Amodei et al., 2016), speech separation (Pariente et al., 2020) or speech synthesis (Oord et al., 2016). Similarly from images, machine learning has allowed for face automatic recognition (Tolba et al., 2006; Valentin et al., 1994) and synthesis (Blanz and Vetter, 1999). Moreover, it has been succesfully used for action recognition from human motion (Lu and Little, 2006; Dalal and Triggs, 2005; Bobick and Davis, 2001). In particular, it has provided crucial contributions to SL automatic processing tasks, such as automatic SL recognition (see Section 3.3.3 or Koller (2020) for a survey).

Beyond automatic recognition purposes, machine learning, jointly with classic signal processing methods, has also allowed gaining insights into the complex structure of human motion. Whether based on video recordings or on 3D mocap data (from simple motion trackers to state-of-the-art full-body setups), several analyses have been conducted in order to quantitatively define properties of human and Sign Language motion (Section 4.1). Some studies have also further explored how individual characteristics may be encoded in motion patterns, for movements such as walking or dancing (Section 4.2). In particular, machine learning techniques have allowed automatically identifying individuals from their movements and, thus, extracting identity-specific features from motion data. Prospects for the development of similar automatic systems controlling identity-specific features of signers in SL movements, in particular for motion synthesis, are discussed in Section 4.3.

4.1 Analysis of human and Sign Language motion

Human motion is governed by biomechanical constraints and motor control laws. In addition to the rules followed by human motion, SL motion is also governed by linguistic rules. For both SL and non-SL movements, specific properties and laws have been established thanks to motion analysis studies. In this section, we elaborate on these properties based on studies using classic signal processing techniques and on others using automatic, machine learning, methods. First, like vocal sounds whose

production and perception is limited in frequencies by human physiological properties, human movements have been shown to lie within a limited frequency range (Section 4.1.1). Within this bandwidth, the motor control system follows specific laws (Section 4.1.2), which can be different when producing SL movements (Section 4.1.3). Moreover, the use of machine learning techniques has allowed providing further data-driven descriptions of motor strategies, including for the identification of pathological movements (Section 4.1.4) and for the automatic evaluation of expertise in sports and musical gestures (Section 4.1.5). We elaborate on these methods, which could be of particular interest for the problem of automatic signer identification. Indeed, if machine learning can be used to automatically classify groups of individuals (e.g., healthy *vs* non-healthy patients, or expert *vs* novice athletes), what about identifying individual signers based on their identity? The underlying problem is highly similar in both cases, as the variable to predict (e.g., expertise or identity) varies from one individual to the other while being invariant across various movement executions. For instance, the predicted level of expertise of a diver should be the same across different dives and the identity of a signer may not vary from one discourse to the other. By contrast, machine learning models used for gesture recognition (e.g., for SL automatic recognition) extract aspects of the movements that are characteristic of the gesture, regardless of the individual who produced it.

4.1.1 Spectral analyses for determining kinematic bandwidths

In order to design relevant models of SL, it is necessary to properly estimate the frequency content of SL movements, provided by motion capture (mocap). First, considering the Nyquist-Shannon theorem (Shannon, 1949; Nyquist, 1928), the sampling rate of mocap systems should be at least twice the actual motion bandwidth. For instance, in the auditory domain, audio signals are often sampled at 44,100 Hz as they have frequencies within the range of roughly 20 to 20,000 Hz, which corresponds to the lower and upper limits of human hearing, respectively. Moreover, state-of-the-art mocap systems now allow for the recording of human movements at high frame rates (e.g., 120, 250 frames per second (fps)). With such high frame rates, mocap data may be noisy and thus are often filtered for human motion analyses (Zago et al., 2017a; Carlson et al., 2020), which requires estimating an optimal cutoff frequency. Up to now, it is unclear what actual bandwidth should be taken to properly model human motion, in particular SL motion. SL movements differ from non-linguistic ones, as they are constrained by not only biomechanic but also linguistic rules. More specifically for technological application perspectives, this problem must be answered in order to better understand whether the spectral content of SL motion is entirely represented when extracted from videos at low frame rates (e.g., 24 fps) (Cao et al., 2019). Indeed theoretically, videos sampled at 24 fps convey spectral information up to 12 Hz only.

The estimated bandwidth of human arm and head motion lies between 2 and 20 Hz, according to Bishop et al. (2001). While investigating gait kinematics, Winter (2009) has reported that most of the spectral energy of a walking body was in a 0–6-Hz range. His analysis revealed more rapid movements for markers on the foot (e.g., heel or ball), which produced frequencies up to 6 Hz, and slower movements for markers on the upper body (e.g., hips or ribs), which produced frequencies up to 3 Hz only. More recently, Skogstad et al. (2013) also have shown that rapid arbitrary motion of the hand may have an upper-bound frequency between 15 and 20 Hz. As shown in Figure 4.1, these rapid hand movements had significant power up to approximately 20 Hz.

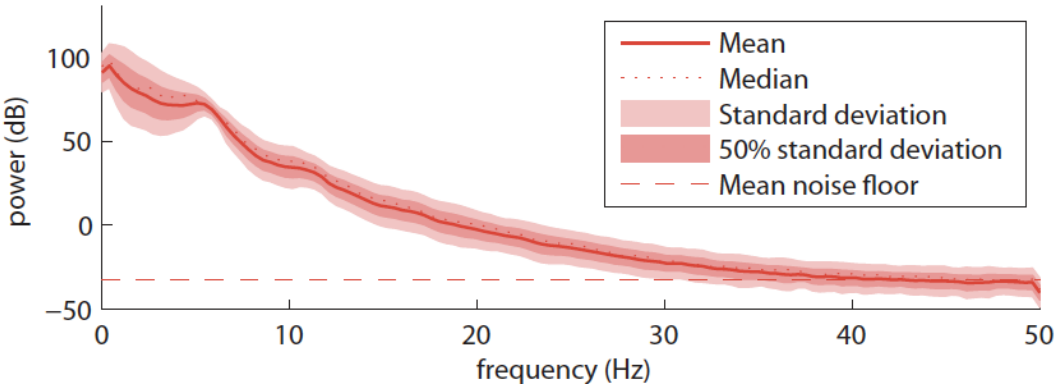


FIGURE 4.1: Spectral power estimation of rapid arbitrary hand movements in Skogstad et al. (2013). The data presented were averaged across 20 mocap recordings. The dashed horizontal line represents the estimated noise level of the mocap recordings.

The spectrum of SL motion has been investigated with isolated signs. Individual lexical signs were produced by one signer and taken out of context. Poizner et al. (1986) have suggested that most of the energy of SL motion may lie below 6 or 7 Hz. According to Foulds (2004), a 0–3-Hz range is enough to understand American SL (ASL) isolated signs and fingerspelling. In the latter study, a first frequency estimation was carried out on the movements of the dominant-hand index finger of an ASL signer producing isolated signs, by means of an electromagnetic position and orientation sensors (see examples of the spectral analyses in Figure 4.2). These analyses suggested that the spectral energy of the signed movements was predominantly below 3 Hz. A further visual perception experiment was then conducted during which human participants were asked to identify various ASL signs from stick figures similar to Point-Light Displays (PLDs), at different levels of spatial or temporal compression. Although spatial compression significantly decreased the accuracy of participants to identify the signs, limiting the spectral range of the stimuli to 0–3 Hz had no impact on the intelligibility of the signs. Sperling et al. (1985) also have reported no significant intelligibility loss for ASL isolated signs from 30 to 10 fps, suggesting a 0–5-Hz bandwidth. In the present thesis, the extent to which these results shown for isolated signers are confirmed for more realistic, spontaneous, SL productions was assessed (see Chapter 6). In McDonald et al. (2016), a signal processing tool was developed for properly removing the noise from continuous SL mocap data. For that aim, McDonald et al. (2016) used estimates of the frequency content relevant for SL modeling. Therefore in Chapter 6, we discuss our quantitative results as compared to their estimation, which interestingly support similar conclusions about the frequency ranges to consider when modeling continuous SL motion for animation purposes.

4.1.2 Human motor control and laws of motion

Human motion is multi-segmental and involves a high number of degrees of freedom (Bernstein, 1966; Saltzman, 1979). Individuals thus have to master the resulting high-dimensional space of potential solutions to successfully execute a particular movement (Hebb, 1949). This high-dimensional problem is achieved thanks to the human motor control system which observes internal and external constraints in order to make the movement successful. Analyses of human movements have allowed

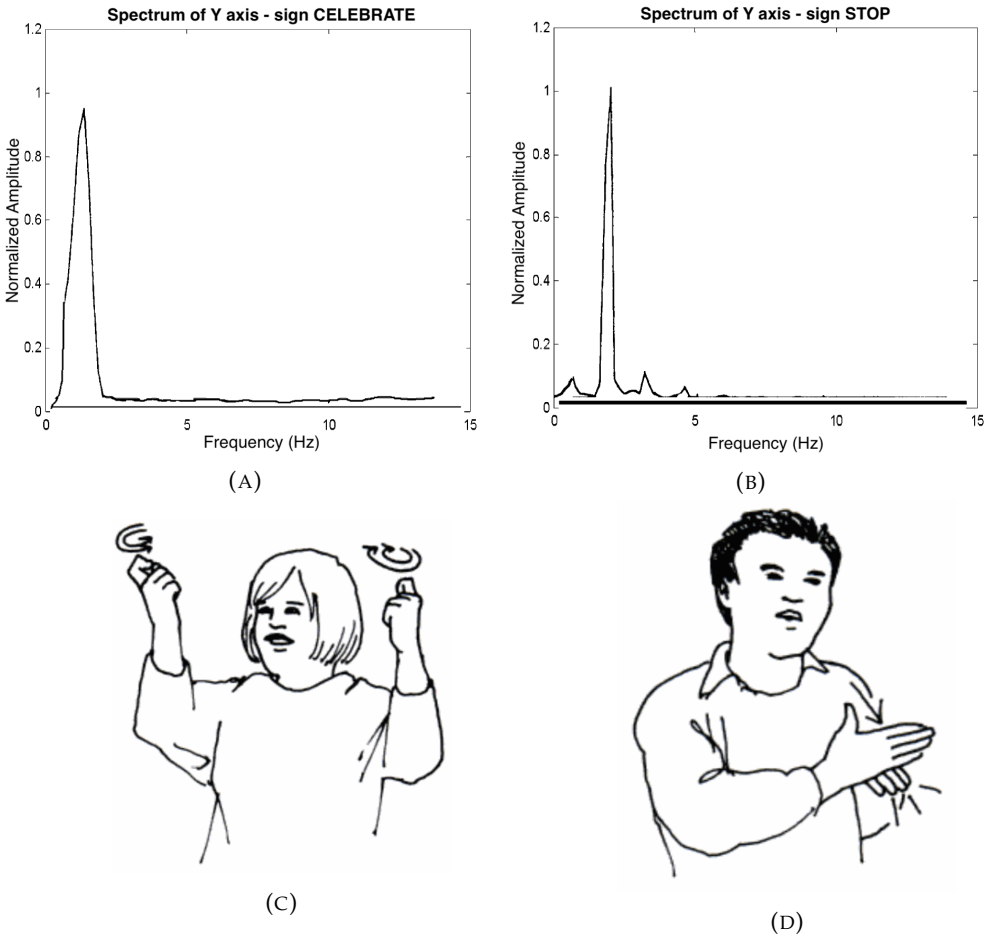


FIGURE 4.2: Spectra of the movements of the dominant-hand index finger along the Y axis in Foulds (2004) for two ASL signs: (A) sign CELEBRATE, made by circling the hands in the air (C), (B) sign STOP, made with a chopping action of the dominant hand into the palm of the nondominant hand (D) (Sign picture descriptions taken from Tennant et al. (1998)).

shedding light on specific motion principles that may be the reflection of general rules followed by the human motor control system in order to solve this problem.

For instance, Viviani and McCollum (1983) have shown that the speed of drawing movements was increased proportionately with the trajectory distance, in order to keep the execution time of these complex trajectories independent of the movement size. The study of kinematics during reach-to-grasp movements of macaques has also outlined that the amplitude of arm peak velocity was correlated with the distance to be covered (Sartori et al., 2013). Termed as isochrony principle, this linear relationship between velocity and the extent of the trajectory has been demonstrated for a variety of actions, such as writing (Michel, 1971), lifting weights (Gachoud et al., 1983), kicking activity in infants (Thelen and Fisher, 1983) or hand and arm movements (Freund and Büdingen, 1978). Further studies have shown that the isochrony principle was independent of age for drawing movements of 5- to 9-year-old children, while context had a significant impact on the strength of this relation between speed and trace length (Vinter and Mounoud, 1991). Moreover, the isochrony principle can be affected by the range of the movements. For instance,

when executing wide movements, although velocity is increased, the movement duration may also increase (Berret and Jean, 2016). In other words, although the motor control system compensates the increase in amplitude by increasing the movement velocity, this phenomenon is not always sufficient to ensure the independence of duration from the movement extent.

In the case of curvilinear trajectories, the kinematic analysis of drawing movements has revealed another well-known law of motion: the two-thirds power law (also called one-third power law) (Lacquaniti et al., 1983). The results of the latter study have suggested that the velocity of the drawing movements was correlated to the value of the curvature, so that velocity increases in less curved portions of the trajectory and, inversely, decreases in more curved portions of the trajectory. The two-thirds power law was then formulated in the following ways:

$$A(t) = kC(t)^{\frac{2}{3}} \tag{4.1}$$

$$\log A(t) = \log k + \frac{2}{3} \log C(t) \tag{4.2}$$

where $V(t)$ is the tangential velocity, $A(t) = V(t)/R(t)$ is the instantaneous angular velocity, $R(t)$ is the radius of curvature, $C(t) = 1/R(t)$ is the curvature of the trajectory and k is the gain velocity factor, which depends on the tempo of the movement and is often considered as constant by the total length of the trajectory or within units of motor action.

These equations can also be formulated in an equivalent way, which justifies the other designation of the law (i.e., one-third power law):

$$V(t) = kR(t)^{\frac{1}{3}} \tag{4.3}$$

$$\log V(t) = \log k + \frac{1}{3} \log R(t) \tag{4.4}$$

Equations 4.3 and 4.4 are illustrated in Figure 4.3, which demonstrates the two-thirds power law (i.e., the linear relationship between the logarithms of the tangential velocity and the radius of curvature for elliptical movements).

In addition to the 2D drawing movements investigated in Viviani and Schneider (1991), the two-thirds power law has been confirmed in further cases, such as 3D drawing (Massey et al., 1992) or eye movements (de'Sperati and Viviani, 1997). It also has been used as an evidence that certain properties of the motor system implicitly influence perceptual interpretation of the visual stimulus (see perception-action coupling theory in Section 2.1.2) (Viviani and Stucchi, 1992). Moreover, if age had no effect on the isochrony principle in drawing movements of 5- to 9-year-old children (Vinter and Mounoud, 1991), Viviani and Schneider (1991) have outlined age-dependent differences for both the phenomena of isochrony and two-thirds power law between the movements of adults and children. In the following sections, we further elaborate on how human attributes, such as age but also gender or identity, could be automatically identified by specific aspects of motion using machine learning techniques.

Other important laws of motion have been formulated thanks to motion analysis. For instance, individuals need to reduce the speed of their movements in order to ensure the accuracy of their action (i.e., "speed-accuracy trade-off"). This phenomenon, now called Fitt's law, has been formalized by Fitts (1954) through an equation linking the time required to rapidly move to a target area with the distance

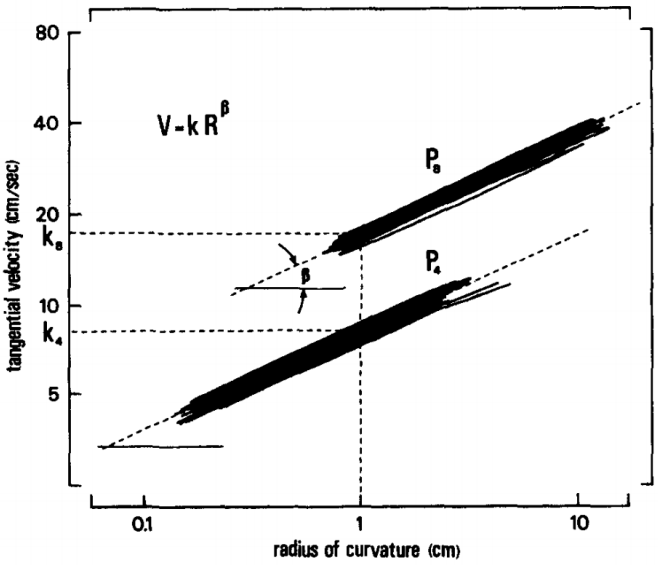


FIGURE 4.3: Illustration of the two-thirds power law from Viviani and Schneider (1991): linear relationship between the logarithms of the radius of curvature and the velocity of the hand during the drawing of ellipses of perimeters 6.63 cm (P4) and 26.51 cm (P8). Following experimental measurements, the β exponent in the authors' formula is a constant that takes values close to 1/3.

to the target and the width of the target. Moreover for vertical arm movements, it has been shown that the motor control system optimizes the energy expenditure, by integrating gravitational forces and thus processing an upward movement differently than a downward movement (Berret et al., 2008).

Beyond the effects of age and context mentioned above, recent research has questioned the impact of the linguistic structure of SL movements on laws of motion, investigating the isochrony principle, the two-thirds power law and the integration of gravitational forces during vertical movements (Benchiheub et al., 2016b; Benchiheub, 2017). Although the validity of all the three laws was confirmed during SL movements, whether within a sign or between two signs (i.e., transitional movements), the analyses have outlined some intriguing effects of SL on the aspects of both vertical movement law and isochrony principle. For instance, both these laws were unchanged during transitional movements while SL-specific aspects, such as longer acceleration times for vertical movements and stronger emergence of the isochrony principle, appeared during the production of signs. These results suggest that the linguistic property of SL has a significant impact on how the motor control system plans the execution of a signer's movements in SL. In relation to the present thesis, this impact of SL linguistic structure on motor control calls for further investigations on how the extraction of human characteristics from motion could be specific to SL, compared to prior research carried out on walking or dancing movements.

4.1.3 The specific parameters of Sign Language movements

Some motion analyses have tackled further intriguing questions related to the specific features in SL movements. As already briefly mentioned in Section 4.1.1 for spectral analyses of motion, SL movements can be executed with varying velocities depending on linguistic properties. For instance, lexical signs have been shown to

be executed faster when they are produced in context, compared to when they are produced in isolation (Braffort et al., 2011). Moreover, Benchiheub et al. (2016b) have shown that not only may the speed of SL productions increase in spontaneous discourses, by contrast with isolated signs, but it may also be affected by the nature of linguistic structures within the discourse. For instance, across the sponteanous SL discourses of four different signers describing pictures in French Sign Language (LSF), the mean velocity of the signer's dominant hand was lower in lexical signs than in depicting signs (e.g., that describe size and shapes of entities), which was lower than in transitional movements between signs. These results suggest that motor control is impacted when movements convey some meaning (e.g., lexical and depicting signs), which may result in slower executions than for non-SL movements. This is in line with SL-specific observations made on motor laws (Benchiheub, 2017) (see Section 4.1.2), which were impacted during the production of signs but not during transitional movements. Benchiheub et al. (2016b) have also shown that the dominant hand of signers could be automatically detected, by comparing the distances covered by the two hands, which is higher for the dominant hand. This observation is the same as with any other human movement: signers preferably use their dominant hand, as it provides faster or more precise performance. Interestingly, the distance ratio between the two hands was quite consistent across different signers.

Motion analysis of American Sign Language (ASL) has also shown that the kinematic properties of linguistic stress can be characterized as peak velocities during the production of a sign (Wilbur, 1999), jointly with larger movements and signs being made higher in the signing space (Wilbur and Schick, 1987). Moreover, in American and Croatian Sign Language, Malaia and Wilbur (2012) and Malaia et al. (2013) have shown that motion kinematics were recruited by signers to express linguistic properties in verb sign production. The productions of telic and atelic verbs were compared. Telic verbs describe events as homogeneous (e.g., *swim* or *walk*) while atelic ones describe events as heterogeneous phenomena involving a change (e.g., *fall* or *break*). In both studies, kinematic features (e.g., peak speed, instantaneous acceleration) of verb signs were affected both by predicate type (telic/atelic) and by the position of the sign within the sentence (medial/final). In LSF, Catteau et al. (2016) have outlined kinematic strategies of deaf poets (e.g., acceleration peaks of the whole-body joints) to convey prosodic variation. Signers may also use their gestures in order to control intonation during their discourses, including movements of the arms, hands and upper-body but also facial expressions (Weast, 2008).

Taken together, these analyses of SL motion suggest that kinematics may convey relevant information about semantic, syntactic and prosodic features. It can be hypothesized that motion features, including kinematics, could also reflect non-linguistic properties of the SL productions of signers, such as identity. Some effects of individual characteristics, such as signer's age, on SL production have been shown. Indeed, the study of LSF mocap from elderly signers has suggested that specific kinematics, such as signing rate, may provide a prosodic characterization for the age of a signer (Blondel et al., 2019). However, to the author's knowledge, there were virtually no attempts to characterize the motion features that allow for inferring the identity of a signer from his or her movements in SL.

Furthermore, beyond the specific problem of signer identification, the present thesis aims to conduct quantitative motion analyses in order to provide novel descriptions of SL movements. All the analyses presented in these first sections (Section 4.1.1, Section 4.1.2 and the present section) have used pre-defined motion variables, including kinematic ones, in order to gain insights into the motion properties

of both human and SL motion. The present thesis aims to propose holistic motion descriptions using data-driven approaches, which may provide unexpected insights into how the complex movements of SL are structured. We thus further discuss how data-driven analyses of human motion have contributed to the understanding of motor control, including postural control, and how it can be useful to quantify specific aspects of a movement for various purposes, such as clinical research or motor learning.

4.1.4 Data-driven approaches for the analysis of postural control and for the identification of pathological movements

As described in Section 4.1.2, the motor control system integrates various internal and external constraints in order to coordinate the multiple segments of the human body and to execute movements in an optimal manner. Pre-defined motion variables, such as instantaneous velocity or acceleration of body joints, have allowed researchers to formalize various laws of motion. As previously introduced in Section 3.2.4, unlike the use of pre-defined variables, some studies have used data-driven approaches to provide holistic descriptions of the movements. In particular, the application of Principal Component Analysis (PCA) to motion data (Federolf et al., 2014) has allowed unveiling human movement strategies for postural control as well as detecting differences in movements between healthy individuals and individuals with pathological conditions.

Some first studies have applied PCA to motion data in order to decompose human movements, such as walking (Troje, 2002a) or skiing (Federolf et al., 2014), into Principal Movements (PMs) (see Section 3.2.4). PM decomposition can be used to quantify the motion patterns utilized by individuals to maintain an upright posture (Federolf et al., 2013b) or to control their posture while executing reaching motor commands (Berret et al., 2009). In addition to the first formulation of the PMs, Federolf (2016) has proposed a further approach deriving "principal positions", "principal velocities" and "principal accelerations". Using a 3D mocap system, the postural movements of participants were recorded when standing on a force plate. The resulting PMs explained the variance of the body Center of Pressure (CoP) with high accuracy. More importantly, these variables have allowed gaining insights into how the postural control system govern the body movements. In particular, the integration of "principal accelerations" has provided information about the neuromuscular control, as muscle synergies are intrinsically linked with accelerations between body segments. Longo et al. (2019) then have also supported that "principal accelerations" obtained by PM decomposition, beyond "principal positions", provide a significant contribution to the understanding of human movement control, in a bimanual repetitive tapping task. Further frequency analyses have shown that postural control involves small-amplitude but accelerated fast movements, which are precisely well captured by "principal accelerations" (Promsri and Federolf, 2020). Based on these observations, the latter study has outlined the relevance of using cut-off frequencies between 5 and 10 Hz when filtering human motion data for the analysis of postural control. Compared to the spectral estimates of human movements presented in Section 4.1.1, this bandwidth is specific to the study of postural control and to the measurement of small-amplitude, but fast, postural movements.

PMs have also provided crucial contributions to the identification of pathological motion patterns. For instance in Federolf et al. (2013a), a 3D mocap setup recorded full-body gait movements of healthy participants and of participants with medial knee osteoarthritis (OA). Group differences were found in the temporal weightings

of the PMs, revealing that greater upper-body (e.g., shoulder or pelvis) movements in OA patients were linked with changes in ground reaction forces, which thus suggests specific compensatory movements of OA patients due to their pathology. Similarly, decomposing human movements into PMs has allowed detecting key aspects of the gait kinematics of children with spastic diplegia (Zago et al., 2017c). 3D mocap recordings of gait cycles executed by healthy children and by children with spastic diplegia were decomposed into their PMs, which allowed detecting changes in the synergies of body segments for pathological movements. For instance, more PMs were needed to explain the gait movements of children with spastic diplegia than those of healthy children. Moreover, these higher-order PMs have then been found to represent compensatory patterns that occured with a high level of individual specificity for the children with spastic diplegia.

In addition to the studies mentioned above, PM decomposition has allowed describing further aspects of human motor control, including age effects. For instance in Haid et al. (2018), PMs were obtained applying PCA to the full-body 3D mocap recordings of young and older individuals performing 80-second tandem stances. Age effects were found only on specific motion patterns (i.e., PM2, PM8 and PM9), older adults presenting less tight and more irregular control in PM2 but tighter control in PM8 and PM9. The extent to which PM decomposition can be used to describe SL motion and to investigate our specific problem of signer identification will be discussed later in this thesis (see Chapters 7 and 9).

4.1.5 Automatic evaluation of gesture expertise

Beyond the study of human motor control, Principal Movements have been widely used to evaluate the degree of expertise of human gestures. The automatic evaluation of gesture expertise is of particular interest for learning complex movements and for improving their execution, using visual feedback of the gestures from poor to excellent performance. Federolf et al. (2014) has been one of the first to apply PM decomposition to sport science. As previously described in Section 3.2.4, the PMs extracted in this study have allowed visualizing the main motion patterns of skiing and, more importantly, comparing the techniques of different athletes. For instance in the first PM, which described lateral body inclination (see Figure 3.9, first row), athletes displayed differences in how fast they tilted their body during the turn and how fast they came back to the upright posture. Visualizing the PMs thus allow for some interpretation of differences in the techniques used by athletes with different levels of expertise. Some studies have further used these descriptions obtained with PMs to train machine learning models in order to automatically evaluate gesture expertise. As the principle of PM decomposition has already been introduced in Section 3.2.4, the present section will mainly focus on how the level of gesture expertise can be automatically predicted from the PMs.

For instance in karate (Zago et al., 2017a), PM decomposition has allowed obtaining five PMs that described most of the information about the movements of both professional and amateur karateka. Then, a linear classifier (i.e., PCA followed by linear regression) successfully predicted the karateka's years of practice from the PMs (i.e., their posture vectors and temporal weightings), jointly with the Center of Mass (CoM) positions and kinematics. First, by comparing the accuracy of the predictions of the model when trained on different subsets of PMs, one can extract the PMs of importance for the evaluation of expertise. Second, by visualizing these PMs of importance for athletes with different levels of expertise, one can quantify the motion patterns that may account for experience level. For instance in Zago et

al. (2017a), experienced karateka raised their left leg higher than amateur karateka during kicking (see Figure 4.4).

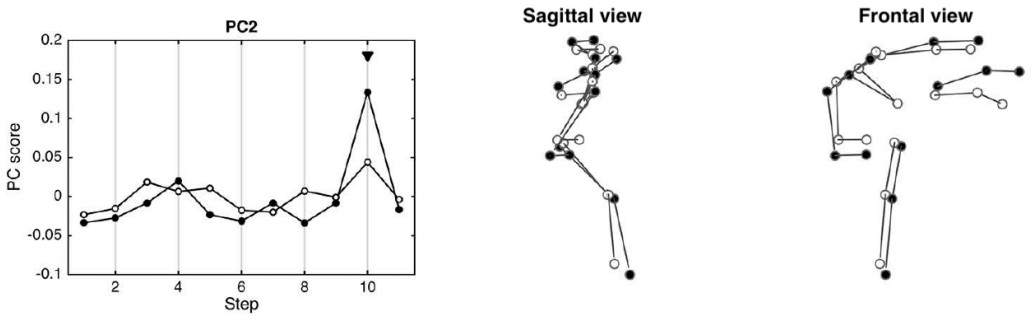


FIGURE 4.4: Example of one PM of importance (PM2) in the model of Zago et al. (2017a) for automatic gesture evaluation. The PM is executed by one karateka with 33 years of experience (black markers) and by one karateka with 5 years of experience (white markers).

Similarly, in addition to classic features used for judging dives (i.e., body center coordinates and splash area), PM vectors and temporal weightings have allowed Young and Reinkensmeyer (2014) to accurately predict actual judges' scores. Like Zago et al. (2017a), the scores were predicted using linear regression. Moreover, novel dives reflecting specific judges' scores were synthesized as stick figures. This method has allowed visualizing how divers could modify their diving performance in order to improve the judges' score.

PMs have also been successfully used to automatically identify motion patterns related to experience in juggling (Zago et al., 2017b). As shown in Figure 4.5, although the dimensionality of the PMs was the same between the two groups in juggling up to 4 balls, intermediate jugglers showed higher-order PMs compared with experts in the most complex task (i.e., 5-balls juggling). Most of these higher-order PMs reflected upper limbs movements, which may be unwanted movements used by intermediate jugglers to compensate for throwing errors. In other words, the reduction of higher-order PMs resulting from years of practice may reflect how postural movements are adjusted by the motor system in order to facilitate the juggling performance.

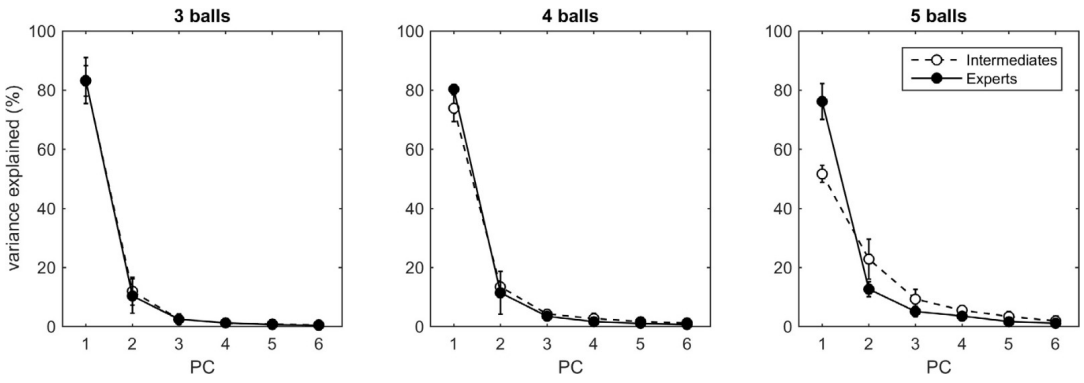


FIGURE 4.5: Automatic identification of differences related to experience in juggling (Zago et al., 2017b). For complex tasks (i.e., 5-balls juggling), movements of experts involve less PMs, which outlines the optimized movement synergies of experienced jugglers.

In the artistic domain, Tits et al. (2015) have extracted PMs from finger gestures

in piano playing using a 3D mocap system, which recorded movements of the articulations of the hands of four pianists. Across different piano pieces, the finger movements of more trained pianists involved more PMs (i.e., eight to ten PMs) than those of less trained pianists (i.e., six PMs). Interestingly, the relationship between the number of PMs involved in the gestures and the level of gesture expertise is in line with the prior investigations mentioned above for juggling (Zago et al., 2017b). However, it yields different interpretations. While higher-order PMs used by intermediate jugglers in Zago et al. (2017b) seemed to reflect unwanted compensatory movements, the higher number of PMs involved in the finger gestures of expert pianists may be due to the finer movements learnt through more years of practice and acquired by playing pieces of higher complexity.

In most of the studies mentioned in this section, significant inter-individual differences have been outlined in the execution of movements in sports or music performance. Rather than being only specific to each individual, the extracted motion patterns have allowed distinguishing between groups of individuals as a function of their level of gesture expertise. What about other attributes that characterize human beings, such as identity? The following section will focus on how the field of motion analysis have allowed automatically extracting motion features related to human characteristics, in particular identity, beyond the general kinematic properties of human movements reported in the present section.

4.2 Automatic extraction of human attributes from motion

Human observers are able to infer socially relevant information about individuals from their movements. Some visual perception studies have aimed to determine the type of motion information that allows for such recognition, revealing a potential major role of kinematic cues (e.g., for gender classification or person identification) (see Section 2.1.4). Still, the nature of such kinematic cues remains unclear up to now. In addition to human perception measurements, other approaches, such as machine learning, can provide further insights along this line. As previously shown for faces (Section 4.2.1), machine learning models can be trained to describe human movements and to extract motion features that allow for automatic gender classification (Section 4.2.2) and person identification (Section 4.2.3).

4.2.1 In the footsteps of face processing

The studies mentioned in Section 2.1.4 have tested the impact of pre-selected motion features, which were potential candidate features to carry critical information, on human recognition. The features were manipulated in Point-Light Displays (PLDs) and the effects on the recognition accuracy of human observers were assessed (Mather and Murdoch, 1994; Kozlowski and Cutting, 1977; Troje et al., 2005; Westhoff and Troje, 2007). Using machine learning techniques, further studies have been able to treat the problem without any a priori assumption about candidate features. These studies have trained machine learning models (e.g., classifier, linear regression models) to automatically detect characteristics of individuals, such as gender or identity. Then, the discriminant features used by the machine learning model could be scrutinized, in order to gain insights into the critical information contained in the movements for the recognition.

For that aim, first studies have developed linear models (i.e., that optimizes a linear mathematical function between the input variables and the output). For instance, Troje (2002a) has trained a linear classifier for the automatic gender classification of

gait. These models took their inspiration from prior work on faces images. Several studies have applied PCA to pixel-based representations of faces, which provided ideal low-dimensional descriptions of the faces for different purposes. Indeed, as shown by Sirovich and Kirby (1987), eigenvectors (called eigenfaces) can be obtained from the application of PCA to an ensemble of face images (see Figure 4.6). Then, any face can be defined in a coordinate system formed by the eigenfaces. In the latter study, cropped faces could be reconstructed as a linear combination of a reduced set of eigenfaces, as shown in Figure 4.7. Sirovich and Kirby (1987) have shown that the error between the reconstructed face and the original one was reduced to 3% when using 40 eigenfaces. This suggests that a face can be characterized using a low-dimensional representation of faces, instead of the whole pixel-based dataset.

FIGURE 4.6: The first eight eigenfaces obtained from an ensemble of cropped grayscale face images in Sirovich and Kirby (1987). Subfigures must be read from left to right, ending at lower right.

Intriguing findings have then been shown using similar low-dimensional representations of faces. As mentioned above, the first eigenfaces of the PC space (i.e., those with larger eigenvalues) may be optimal to reconstruct faces, as they minimize the error between reconstructed and original faces (Sirovich and Kirby, 1987). The first eigenfaces have also been shown to be optimal for the gender classification of

FIGURE 4.7: Example of one face image of the dataset in Sirovich and Kirby (1987), reconstructed using 10, 20, 30 and 40 eigenfaces. Subfigures must be read from left to right, ending at lower right.

faces. For instance in O'Toole et al. (1993), of 100 eigenfaces, the highest correlations between eigenface weights and gender were found in eigenfaces with largest egeinvalues, most of them being within the 15 first eigenfaces. In particular, the largest amount of explained variance of the gender variable was found for the second eigenface, as shown in Figure 4.8. For instance, adding the second eigenface to the first one generates a masculine face (Figure 4.8(c)), while subtracting it from the first eigenface generates a feminine face (Figure 4.8(d)).

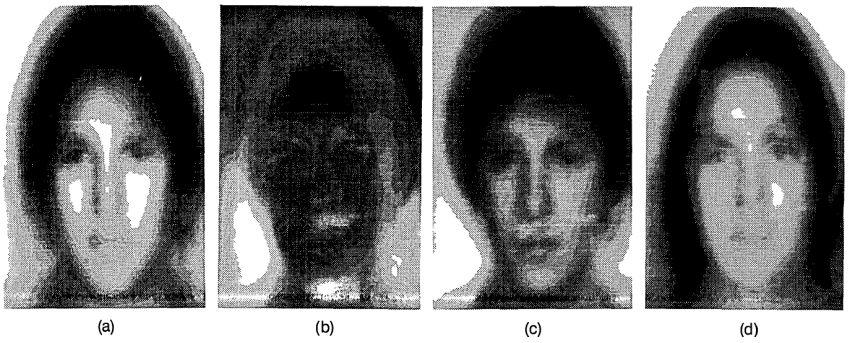


FIGURE 4.8: Low-order eigenfaces are optimal to characterize gender (O'Toole et al., 1993): (a) First eigenface, (b) Second eigenface, (c) Second eigenface added to the first one, (c) Second eigenface subtracted from the first one.

However, low-order eigenfaces may not be optimal for recognizing particular faces. Interestingly, while these low-order eigenfaces are optimal for face reconstruction and for gender classification, higher-order eigenfaces (i.e., with smaller eigenvalues) have been shown to convey much more information for face recognition (O'Toole et al., 1993). A computational model was trained to distinguish between known and unknown faces from reconstructions of faces achieved using different ranges of 15 eigenfaces. Like Sirovich and Kirby (1987), the error between reconstructed and original faces decreased as the range of the used eigenfaces changed from low-order to high-order eigenfaces. However, the ability of the model to recognize the reconstructed faces increased with ranges of higher-order eigenfaces.

Ranges between 45 and 80 eigenfaces allowed for the highest recognition accuracy. This phenomenon is illustrated in Figure 4.9. When reconstructing a face using mainly high-order eigenfaces, information about the identity of the person remains quite visible (Figure 4.9(a), centre and right panels), while when reconstructing the same face with only low-order eigenfaces, the person can hardly be identified (Figure 4.9(b), centre and right panels).

FIGURE 4.9: High-order eigenfaces carry accurate information for face recognition (O'Toole et al., 1993): (a) the original face (left), its reconstruction without using the first 20 eigenfaces (center) and without using the first 40 eigenfaces (right); (b) the original face (left), its reconstruction using only the first 20 eigenfaces (center) and using only the first 40 eigenfaces (right).

Other studies have proposed similar low-dimensional representations of faces, but applying PCA to correspondance-based rather than pixel-based representations of images. Correspondance-based representations describe face images with distinct vectors of shape and texture (Vetter and Troje, 1995; Vetter and Troje, 1997). Similarly, the eigenvectors derived from these face representations have allowed for accurate gender classification using linear classifiers, even with smaller errors than when using pixel-based representations (Troje and Vetter, 1996).

Although other techniques have been succesfully used for face recognition (Guo et al., 2000), the fascinating results presented in this section suggest that PCA can be used to decompose a dataset of face images into uncorrelated eigenfaces and that, depending on the amount of overall variance they explain in the original data, these eigenfaces carry discriminant information about different human characteristics, such as gender or identity. In the present thesis, PCA will be applied to different

motion descriptions, notably to find discriminant information for signer identification. The extent to which the techniques used in the present thesis are in line with the previous observations made on faces will be discussed. We will also question whether identity information could similarly be carried by high-order eigenvectors of SL motion, more than by low-order ones.

4.2.2 From faces to motion: the example of gender classification of gait

As we have already pointed out, the analysis of human motion is complex, notably because of the high number of body articulators and, thus, degrees of freedom involved in the movements. The potential of PCA to reduce dimensionality is thus of particular interest for motion models. As shown for faces in the previous section (O'Toole et al., 1993; Troje and Vetter, 1996; Sirovich and Kirby, 1987), PCA has been applied to motion data in order to obtain uncorrelated motion patterns and to investigate whether these patterns could carry discriminant information related to human characteristics of the moving persons.

For instance, Troje (2002a) has applied PCA to 3D mocap data of human gait. As already mentioned in Section 3.2.4, the first four Principal Movements (PMs) (or eigenmovements) obtained by this PCA explained 98% of the overall variance in the walking movements, across walkers. Compared to eigenfaces, the description of PMs is more complex as they vary over time. However, because of the specific structure of walking movements, Troje (2002a) has been able to describe the temporal behavior of the four PMs with simple sine functions, each characterized by its frequency, its amplitude, and its phase. This decomposition is thus very similar to Fourier decomposition, priorly used in Unuma and Takeuchi (1991) and additionally described in Troje (2002b), where the two methods are compared for the classification of gender or other attributes. However, unlike Fourier decomposition, PM decomposition can be applied to any type of movement including non-periodic ones, which is of particular interest for the analysis of complex SL movements carried out in the present thesis.

PMs are defined by their posture vector (i.e., eigenposture) and their temporal behavior. In Troje (2002a), any gait could thus be reconstructed using the average posture vector, the four eigenposture vectors and the frequency and phase corresponding to their sinusoidal temporal behavior. From these features, a second PCA followed by a linear classifier was applied in order to predict the gender output. The weights of the classifier were optimized using a set of walking patterns (i.e., training data) and their respective labels (i.e., 1 if the walker was a man, -1 if the walker was a woman), giving greater importance to some motion features than to others, depending on how they allow for accurate classification of new gait examples. Using this approach, Troje (2002a) has been able to show that, apart from size, the structural information contained in the average posture of walkers does not play a major role in gender classification. Indeed, the model classified the gender of new gaits with a 15% error when trained on the four size-normalized eigenpostures, their phases and their frequency, while classification error was 17.5% when the model was trained on the full vector, which additionally included the average posture vector.

As a reminder, one motivation of the present thesis is to further determine the aspects of motion, in particular SL motion, that carry critical information about the moving person, in particular identity. We already mentioned that kinematics may play a major role in both person identification and gender classification by human observers (see Section 2.1.4), which is confirmed by the computational results of Troje (2002a) for automatic gender classification of gait. It has also been outlined

that behavioral experiments were quite limited in further quantifying the specific kinematic parameters that allow for these recognitions. We argue that this limitation can be overcome using machine learning models. In Troje (2002a), neither did walking frequencies nor relative phases of walking patterns seem to play a major role in predicting gender. For instance, training the model with the relative phases of the four PMs led to a 37.5% classification error. Similarly, this machine learning approach has allowed quantifying the distinct roles of the four PMs in predicting the gender of walkers. For instance, training the model with the first PM alone resulted in a performance almost as good as the one obtained with all four PMs.

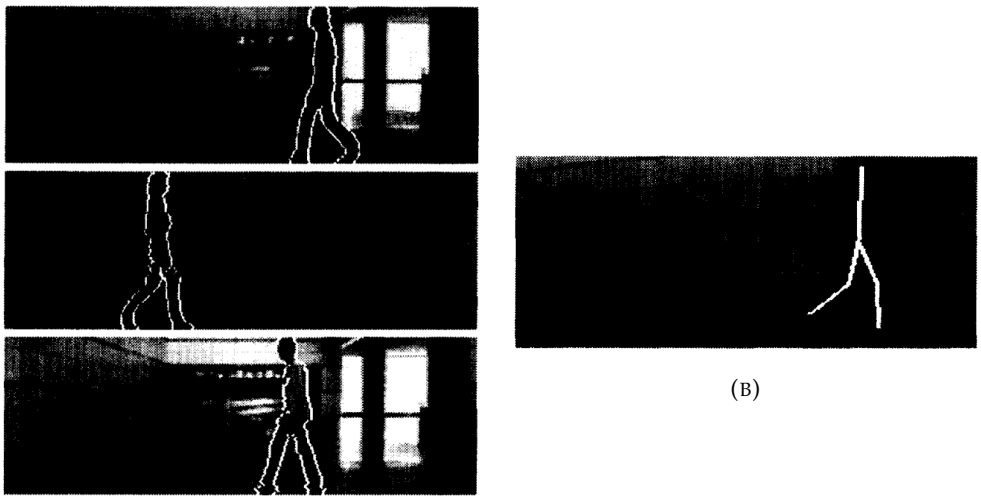
In summary, using the approach described above, the importance of various input motion features in classifying gender can be scrutinized. Note that the model of Young and Reinkensmeyer (2014) for automatic gesture evaluation of dives (see Section 4.1.5) was highly similar to the dual-layer PCA approach of Troje (2002a) presented in this section, except that the first model was trained to predict a linear variable (i.e., quality of dives) while the second was trained to predict a binary variable (i.e., gender). Machine learning models thus have opened up new possibilities to unveil the encoding of relevant information, such as expertise or gender, in human movements. What about identity information?

4.2.3 Person identification from motion: from gait to SL movements

Following the same procedure as in Troje (2002a), Zhang and Troje (2005) have been able to automatically identify walkers from 3D mocap data. PCA was applied to key postures (analogous to the eigenpostures used in Troje (2002a)) across various walkers and the resulting principal components allowed predicting the identity of the walker from different viewpoints. Although the methodology was highly similar to the one used for gender classification, Zhang and Troje (2005) have put a further focus on the impact of viewpoint on automatic person identification. Overall, all viewpoints allowed for high identification accuracy ($> 91.8\%$). However, the highest performance of their model (98.8%) was obtained from a 3/4 view, which is in line with prior work on face recognition (Bruce et al., 1987).

One of the first studies to develop algorithms for the automatic identification of moving persons may be Niyogi, Adelson, et al. (1994). From gait video sequences, the contours of five different walkers were extracted and were used to define stick figure models. Applying a simple nearest neighbor technique with Euclidean distances to the trajectories of the stick model (two sticks per leg, and another for the torso, see Figure 4.10), their identification methods reached recognition rates as high as 81%. Later, Little and Boyd (1998) have shown that automatic identification was possible using optical flow in gait videos. Time series of moving points were obtained from optical flow images. Across walkers, these series shared similar walking frequency but differed in their respective phases. Automatic identification was thus accomplished using the phase vectors as input. Similarly to Niyogi, Adelson, et al. (1994), the walker was predicted taking the nearest neighbor of its gait sequence, in terms of Euclidean distance between phase vectors. Using this approach, the accuracy of identification could reach 95.2%. Interestingly, the latter analysis has aimed to gain insights into the features that allow for person identification, rather than to focus on identification performance *per se*, which is a central question of the present thesis. In that regard, the reported minor role of walking frequency in automatic identification is in line with behavioral data (Troje et al., 2005).

Another intriguing approach is the one taken by Murase and Sakai (1996), which additionally emphasizes how PCA-based face recognition algorithms have



(A) (B)

FIGURE 4.10: Gait data extracted from 2D videos for walker identification in Niyogi, Adelson, et al. (1994). (A) walker contours, (B) stick figure model.

been a source of inspiration for person identification from motion. Moving silhouettes of walkers were extracted from video images and then decomposed into eigenvectors that captured the spatio-temporal changes of the silhouette. As observed by Troje (2002a) and Troje (2002b) (see Section 4.2.2), eigenvectors of gait were intimately linked with walking frequencies (e.g., low-order eigenvectors corresponded to low frequencies while higher-order eigenvectors corresponded to higher frequencies). Recognition from the projections of gait sequences onto this eigenspace allowed identifying 10 different walkers with accuracy levels up to 100%. Egeinspace transformation, jointly with other motion descriptors, has similarly allowed for accurate automatic identification of walkers from videos in Huang et al. (1999).

PCA-based and Fourier decompositions have allowed modeling gait patterns and successfully training machine learning systems to identify walkers. Using these methods, some studies have been able to provide insights into the specific roles of gait parameters in identifying the walkers. For instance, some studies mentioned above have reported high identification accuracy using phase information (Little and Boyd, 1998). Moreover, other analyses precisely have pointed out the important role of phase information in walker identification, revealing that walkers were better identified when using the magnitude spectra of gait patterns multiplied by its respective phase, rather than when using the magnitude spectra alone (Cunado et al., 1997; Cunado et al., 2003). These observations can be surprising, as phase information did not seem to play a major role in automatic gender classification of gait (Troje, 2002a). This illustrates how the encoding of identity in motion patterns is still unclear and how computational, including machine learning, analyses of motion can contribute to further determining the motion features that carry identity.

In more recent studies, the discriminant representations provided by deep neural networks have allowed automatically identifying walkers from their gait (Babaee et al., 2019; Huynh-The et al., 2020; Liu et al., 2018) or "re-identifying" them (i.e., re-associating a specific walker across non-overlapping cameras) (Yan et al., 2016), using RGB and depth videos. However, these features are often hard to interpret. Moreover, all of the identification methods presented above have been applied to gait, which is very specific. Gait patterns present a temporal behavior quite easy to model, in particular using sine functions. That is hardly the case of SL movements,

which involve a wide variety of motion patterns, from various body parts and within a continuous and complex stream.

All of the methods mentioned so far in this chapter form the theoretical foundations of the computational models that will be developed in the present thesis, but some recent studies have been particularly inspiring in how machine learning models can extract relevant information from motion data that are neither defined by a clear invariant temporal behavior nor synchronized in time across motion examples and individuals. For instance, a machine learning model has been successfully trained to identify dancers (Carlson et al., 2020) from time-averaged statistics of 3D full-body mocap data. Using a non-linear classifier (Support Vector Machine), the mocap data of 73 individuals dancing freely to music allowed for automatic person identification with a 94.1% accuracy, across eight musical genres. As shown in Figure 4.11, the performance of the model was significantly higher for person identification than for musical genre classification, despite a lower chance level (1.37% compared to 12.5%). This study has demonstrated that the identity of a dancer may be encoded by the covariance of 3D movements between specific body markers.

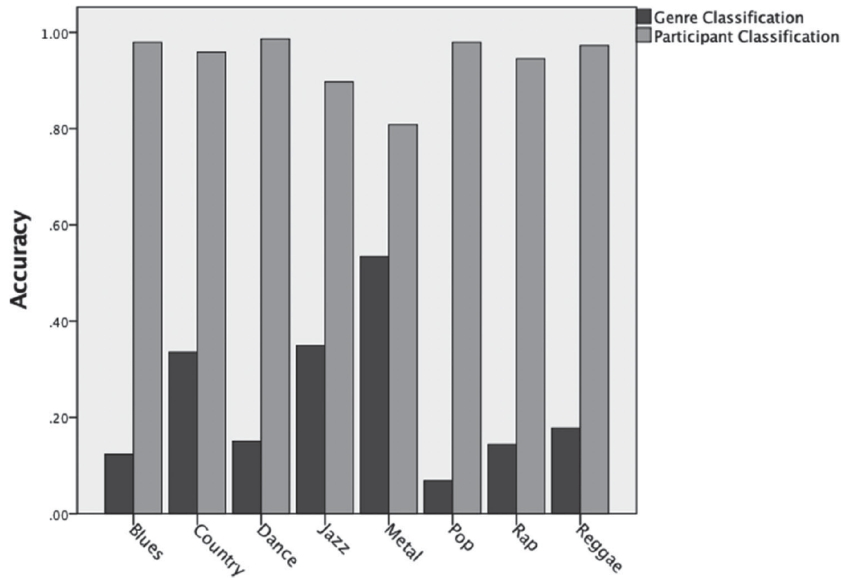


FIGURE 4.11: Accuracy of the model of Carlson et al. (2020) for automatic classification of musical genre and for person identification (termed as "participant classification" in the original study) from dancing mocap data, per musical genre.

Interestingly, the results of Carlson et al. (2020) support that most of the critical information for identifying dancers was conveyed by motion kinematics, as for the automatic person identification (Zhang and Troje, 2005) and gender classification (Troje, 2002a) of walkers. Similarly, for the automatic evaluation of gesture expertise, a linear regression model trained by Tits (2018) has been able to accurately predict the level of expertise from gesture in Taijiquan, based on mean and standard deviation of position and velocity. As a reminder, unlike this statistical-based approach, all the gesture evaluation models presented in Section 4.1.5 relied on temporal, frame-by-frame, comparisons of the different gestures to assign an accurate expertise level. Of all the machine learning models mentioned in this chapter, a summary of those of particular relevance for the present thesis is presented in Table 4.1. The general framework shared by all these models can also be found in Section 4.3, with the additional final step related to motion synthesis and visual feedback.

TABLE 4.1: Summary of important machine learning (ML) models of motion, from various motion representations and for various automatic problems.

Study	Type of motion	Problem	Motion features	Dimension reduction	ML model
Young and Reinkensmeyer (2014)	Diving	Evaluation of expertise	PCA-based: – Eigenpostures – Their temporal curves – The body center 3D position – The splash area vector – The board tip 3D position	PCA	Linear regression
Zago et al. (2017a)	Karate	Evaluation of expertise	PCA-based: – Eigenpostures – Their temporal curves – The CoM 3D position – The CoM 3D velocity	PCA	Linear regression
Troje (2002a)	Gait	Gender classification	PCA-based: – The average posture – Eigenpostures – The fundamental frequency – The relative phases	PCA	Linear regression
Troje (2002b)	Gait	Gender classification	Fourier-based: – The average posture – Key postures (related to harmonics) – The fundamental frequency	PCA	Linear regression
Zhang and Troje (2005)	Gait	Person identification	Fourier-based: – The average posture – Key postures (related to harmonics)	PCA	Nearest neighbor (Euclidean distance)
Carlson et al. (2020)	Dancing	Person identification	Statistical-based: – Covariances of velocity for all body markers in 3D	L1-norm SVM	L2-norm SVM

Taken together, these results call for further investigation into the role of kinematic cues in the perception of an individual's identity, in particular for SL. As already mentioned, one specific aspect of SL is to be governed not only by biomechanic rules, but also by linguistic ones, which may thus reveal SL-specific signatures for signers' identity. The present thesis aims to test both frame-by-frame and time-averaged statistical approaches. Still, as shown for speaker identification in the auditory domain (Latinus and Belin, 2011; McDermott et al., 2013), it is hyposethized that statistics may provide a particularly well-suited description to extract identity information, as identity is a time-invariant property that humans are able to recognize from different utterances of the same individual. Moreover, in addition to the determination of motion features that allow for signer identification, the further objective of this thesis is to manipulate the discriminant features in motion generation models, in order to control the identity of signers in the movements of SL animations.

4.3 From automatic recognition to synthesis

We mentioned several times in this chapter the potentials brought by PCA and PM decomposition (see Section 3.2.4) in terms of dimensionality reduction. One other main advantage of PM decomposition is that it allows resynthesizing PMs in the original 3D space. Motion sequences are generated from a linear combination of the PMs, whose temporal weights can be manipulated in order to either amplify or reduce the impact of specific PMs in the synthesized motion. It has been widely used for evaluating the quality of gestures (see Section 4.1.5), as it provides athletes with a visual feedback highlighting the aspects of specific motion patterns that can be modified in order to improve their performance. In that regard, these methods could also help sports coaches and facilitate motor learning in various domains, such as sports but also music performance.

This technique can be generalized to many other types of motion representations and for the manipulation of many other attributes, such as gender or identity. The general framework can be defined as follows (for further details about the specific models, see Table 4.1):

- First, the motion data are described using a given representation, or feature vector, which often includes various features (e.g., PM posture vectors, their temporal weightings and the CoM positions and kinematics in Zago et al. (2017a), or the average posture vector, PM posture vectors and their respective phase and frequency in Troje (2002a)).
- Very often, the dimensionality of this feature vector is reduced using techniques such as PCA (Zago et al., 2017a; Young and Reinkensmeyer, 2014; Troje, 2002a; Tits, 2018) or L1-norm SVM-based feature selection (Carlson et al., 2020).
- Then, a machine learning model (e.g., linear regression in Zago et al. (2017a), or L2-norm SVM in Carlson et al. (2020)) is trained on the feature vectors of the training data, in order to accurately predict the desired output from the feature vectors of novel data. To do so, the machine learning model optimizes the weights of its mathematical function that links the input variables to the output variable. These classifier weights can then be scrutinized in order to find the motion features that have been used.
- Features that have been assigned a high weight by the classifier are diagnostic for the classification (e.g., gender classification or person identification). Therefore, novel motion sequences can be synthesized by either amplifying or reducing the importance of these features in the movements. A convincing illustration of that concept is shown for faces in Figure 4.9 (O'Toole et al., 1993) where, depending on the weight given to the second eigenface in the reconstruction, the reconstructed face could appear to be feminine or masculine.

The fourth step of the framework presented above raises a main question for the synthesis of moving patterns with controlled attributes: how to amplify (or reduce) the importance of discriminant features in the synthesized movements? The answer to that question is entirely dependent on the motion representation used in the first step. As mentioned in the first paragraph, many studies have used PMs to describe the motion data, which were easily manipulated based on their impact on classification. For that aim, the PM posture vectors and their respective weights were modified in the linear combination used to reconstruct the original data (Young and Reinkensmeyer, 2014). In fact, by contrast with face synthesis, one of the main

difficulties of this procedure is to properly model the temporal behavior of the synthesized movements while manipulating certain aspects related to the attribute of interest. In Troje (2002a), this problem was overcome by modeling the temporal curves of the main PMs using sine functions. Based on the weights that the classifier assigned to the different features (i.e., the average posture, the four PM postures and the phases and frequencies defining their temporal sine curves) in order to accurately predict gender of the walkers, new motion patterns could be synthesized while manipulating the gender attribute (see Figure 4.12).

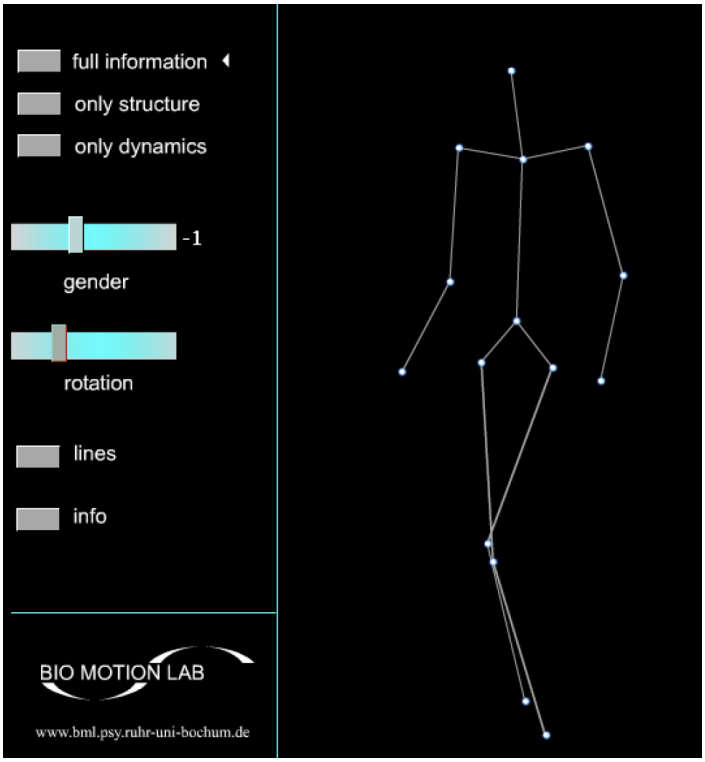


FIGURE 4.12: Interactive application of Troje (2002a) allowing for the synthesis of walking patterns as stick figures, while manipulating the gender attribute. Users can choose whether to impact all the motion features related to gender, or only structural ("only structure") or kinematic ("only dynamics") ones.

Various other methods, using other motion representations, have been applied to generate new motion patterns. For instance, Hidden Markov Models (HMM) (Tilmanne et al., 2012; Tilmanne et al., 2014) as well as Gaussian modeling (Tilmanne and Dutoit, 2010) have been used in order to manipulate the style of synthesized gaits. Both these methods have allowed modeling temporal aspects of the movements, such as the duration of cycles in the gait (Tilmanne and Dutoit, 2010). For gesture evaluation, Tits (2018) has used a Generalized Regression Neural Network (GRNN) based on evaluation scores in order to synthesize motion animations representative of a given skill level. From a motion example executed by the user, the synthesis model aimed to provide a visual feedback showing the user how to improve the quality of his or her movement. In addition to the GRNN, scaling methods, including temporal scaling methods (i.e., Dynamic Time Warping) were used in order to be aligned with the sequence of the user and thus provide consistent feedback.

Therefore, depending on the motion representations that seem best suited for signer identification, one objective of the present thesis is to develop synthesis algorithms that allow manipulating the features that carry identity information in novel

SL motion animations. Based on this theoretical framework, we propose a machine learning model for the automatic identification of signers from motion statistics (see Chapter 9) and present synthesis algorithms for re-generating SL mocap discourses while controlling the discriminant motion statistics that carry identity information, according to the identification model (see Chapter 10).

Part I Summary

From a very young age, humans develop an impressive ability to decode biological movements. For instance, whether from walking, dancing or jumping movements, human observers can recognize human attributes, such as the identity of the moving person. This raises the possibility that signers could be identified from their movements when signing, in the same way as speakers can be identified from their voice when speaking. For now, it is unclear whether Sign Language (SL) users actually manage to identify signers from motion and which parameters of the movements allow for the identification. Determining these parameters and controlling them in the animation of virtual signers (e.g., for anonymization purposes, see Chapter 1) is a complex task, as SL involves a wide variety of motion features. The present thesis aims to develop computational models for the description of these complex movements (Part II) and to uncover the motion information that allows inferring the identity of a signer (Part III). For that aim, models could be used to propose novel relevant descriptors for SL movements and to gain insights into their complex properties, using 3D motion capture data. Furthermore, these models could be tested in order to shed light on the motion features that allow predicting the identity of the signers, as previously done for the recognition of expertise, gender or identity from non-SL movements. The discriminant features could then be controlled in SL animations in order to change the identity perceived by human observers when viewing the movements.

Part II

Kinematic analysis of Sign Language

Chapter 5

MOCAP1: 3D mocap corpus of spontaneous French Sign Language

The 3D SL mocap data used in the present thesis were taken from a previously reported corpus: MOCAP1 (Benchiheub et al., 2016b; Benchiheub et al., 2016a). This corpus provides highly accurate 3D motion recordings of spontaneous French Sign Language (LSF) across multiple signers, which makes it ideally suited for the aims of this thesis (Section 5.1). From the mocap data of the original corpus (Section 5.2), we present some further contributions, including new data representations and methods of preprocessing, normalization and visualization (Section 5.3). Note that, although the following contributions (from Chapter 5 to Chapter 10) were carried out on LSF mocap data, most outcomes can actually be extended to all SLs. When discussing the results, we then often refer to SL (general) or SLs.

5.1 Why MOCAP1?

Up to now, the majority of SL corpora are video corpora (see Section 3.3). RGB 2D videos, jointly with depth videos, have allowed training convincing models for the automatic recognition of SL, notably thanks to the impressive progress of deep learning techniques (see Section 3.3.3). Moreover, recent computational methods have been proposed to extract the trajectories of human skeletons from 2D videos (Cao et al., 2019). Some SL recognition models can even derive the 3D trajectories of these skeletons using deep neural networks trained on both 2D videos and 3D motion data (Belissen et al., 2020). State-of-the-art image processing techniques, including deep learning, could thus enable researchers in SL automatic recognition to be satisfied with video corpora.

For SL motion analysis instead, the lack of 3D mocap corpora can be quite limiting, notably in terms of spatial and temporal precision of the recorded data. For instance, although some techniques mentioned above can be used to estimate the 2D and 3D poses of signers from 2D videos, the estimated trajectories are by definition limited to the spatial and temporal precisions of the video recording, which are significantly lower than those provided by state-of-the-art 3D mocap setups. The trajectories estimates can also be approximate (e.g., inaccurate estimate, missing body joint or confusion between body joints) depending on the performance of the estimation model. Furthermore, most SL video corpora have been created specifically to create lexicons, to develop SL automatic recognition models or to address linguistic problems, rather than to conduct motion analyses, which often makes them not ideally suited for the questions addressed in the present thesis.

At the intersection of multiple disciplines, such as motion science, human visual perception, motion analysis and machine learning, the present thesis required

3D motion data recorded with high accuracy and whose SL content was tailored to our specific problems. For instance, 3D mocap corpora made with only one signer (Malaia et al., 2008; Duarte and Gibet, 2010) hardly allow developing automatic models of person identification, as multiple identity labels are needed to train the model. Furthermore, most existing 3D mocap SL corpora created for motion analyses have been designed very specifically for the purposes of the corresponding study. Therefore, they often provide a limited representation of SL movements in addition to being of short duration. Finally, not only has a limited amount of accurate and well-suited 3D mocap SL data been recorded, but they are often also not publicly available.

MOCAP1 is a 3D mocap SL corpus of spontaneous French Sign Language (LSF) and it has been made fully available (Benchiheub et al., 2016b; Benchiheub et al., 2016a). In brief, the part of MOCAP1 used in the present thesis provides spontaneous LSF descriptions of various pictures produced by multiple signers. The LSF productions are termed *spontaneous* as signers freely described pictures without any constraints in time, signs or structure. Therefore, this corpus allows investigating the complex structure of SL movements in more realistic conditions than other corpora (e.g., ones made in isolation) and how to model it for application purposes (see Chapters 6 and 7). Furthermore, the LSF movements of MOCAP1 have been recorded across different signers, which makes it well suited for the study of person identification. For instance, insights could be gained into the human ability to infer identity from the movements of signers in spontaneous SL discourses (see Chapter 8), into the further specific aspects of motion that carry identity information (see Chapters 8 and 9) and into how to manipulate these identity-specific features in the movements of SL animations, not only for isolated signs but also for real-life interactions (see Chapter 10). To the author's knowledge, the only other 3D mocap SL corpus that could have been used to conduct the present studies is the CUNY ASL corpus (Lu and Huenerfauth, 2014). Testing the models developed in this thesis on other corpora, such as CUNY, could be of great interest in future work, notably in order to assess the extent to which our results would generalize to other motion data and other linguistic contexts.

5.2 Description of the original corpus

This thesis investigated movements of signers taken from picture descriptions in LSF, which are one part of the whole MOCAP1 corpus (Section 5.2.1). Thanks to an optical state-of-the-art mocap setup, this corpus provides 3D motion data of various upper-body parts involved in signing (Section 5.2.2). Extensive description of the corpus collection can be found in the original work of Benchiheub (2017).

5.2.1 Participants and Sign Language discourses

In the original MOCAP1 corpus, eight deaf fluent signers have used French Sign Language (LSF) during five different tasks: (1) description of pictures, (2) translation of short journalistic texts, (3) description of procedures for transport, (4) storytelling based on pictures and (5) storytelling based on movies. The signers were aged from 24 to 58 years old and all of them had high fluency with signing in LSF. Four of them were native signers, two of them learned LSF at the age of 2 and two of them learned LSF later (after 8 years-old). Although being all fluent in LSF, the signers had grown up in different family environments. Half of the signers had grown up with

hearing relatives while the other half had deaf people in their family environment. The dominant hand of all signers was the right hand.

All of the recorded LSF discourses were monologues. Of the five tasks, only the mocap data of the first one (i.e., picture description) have been publicly released (Benchiheub et al., 2016a). The studies carried out in the present thesis were based on this description task. For this task, signers freely described the content of 25 pictures (see three examples in Figure 5.1 and all the 25 pictures in Appendix A). Pictures were chosen to show specific geometric shapes (e.g., rounded, horizontal and vertical forms in Figure 5.1, respectively). After looking at a given picture for a few moments, signers were asked to spontaneously describe it in LSF. Using picture description as a task particularly allowed eliciting spontaneous SL discourses, which are more likely to reflect real-life SL productions. For instance, eliciting SL discourse with fixed sentences (e.g., in French) would constrain the SL syntax, biased by the sentence language (e.g., French) syntax system. Furthermore, picture descriptions elicited a wider variety of SL linguistic forms beyond lexical signs, such as depicting ones. For further details about the four other tasks included in the original MOCAP1 corpus, see Benchiheub (2017).

FIGURE 5.1: Examples of pictures described by the signers in MOCAP1 corpus (Benchiheub, 2017).

5.2.2 Motion capture data

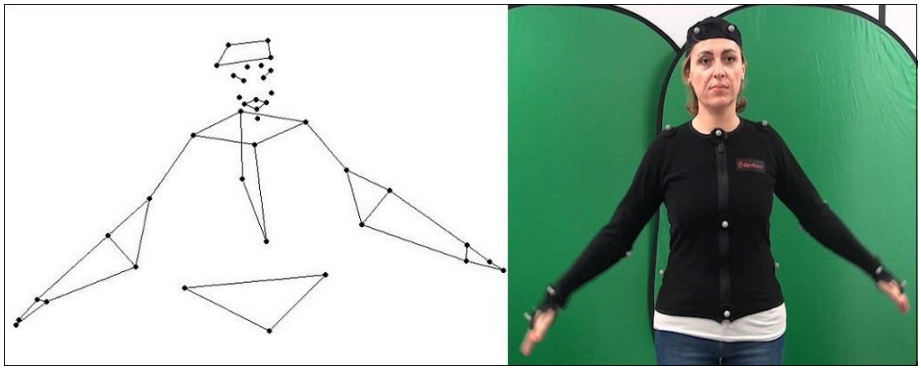


FIGURE 5.2: Arrangement of the markers used in MOCAP1 corpus (Benchiheub, 2017). Left: marker positions seen as mocap data. Right: markers placed on one signer of the corpus.

3D mocap data and 2D video data were recorded during the LSF productions. Using an optical mocap system equipped with 10 cameras (*Optitrack* S250e), the 3D mocap data consisted of the upper-body movements of the signers as well as their

facial expressions, recorded at 250 fps. For that aim, a set of 40 passive retroreflective markers was used. Signers were wearing suits provided with 23 markers to record the movements of upper limbs, including shoulders, elbows, wrists, hands and chest. Four additional markers were attached to a cap in order to record the head movements. A set of 13 markers was used to record the movements of various facial parts, including eyebrows, eyelids, cheeks, chin and mouth. The arrangement of the markers is shown in Figure 5.2. Finger movements were not recorded. All markers were described as 3D global Cartesian coordinates in an external reference system with the center of the room as origin. For further technical details about the recording protocol, such as camera placement, camera settings or calibration, see Benchiheub (2017).

In parallel with 3D mocap data, 2D videos were recorded using a classic HD videocamera sampled at 25 fps. The mocap recordings then could be downsampled from 250 to 25 fps in order to be synchronized with the video data for corpus annotation. Different types of SL movements were manually annotated using ANVIL (Kipp, 2001), notably in order to distinguish signed content from transitional (i.e., between two signed entities) movements. As shown in Figure 5.3, the annotation data provide information about eye gaze, manual signs and movements of the two hands. An extensive description of the annotation data is out of the scope of this thesis, as they were finally not used (for further details, see Benchiheub (2017)).

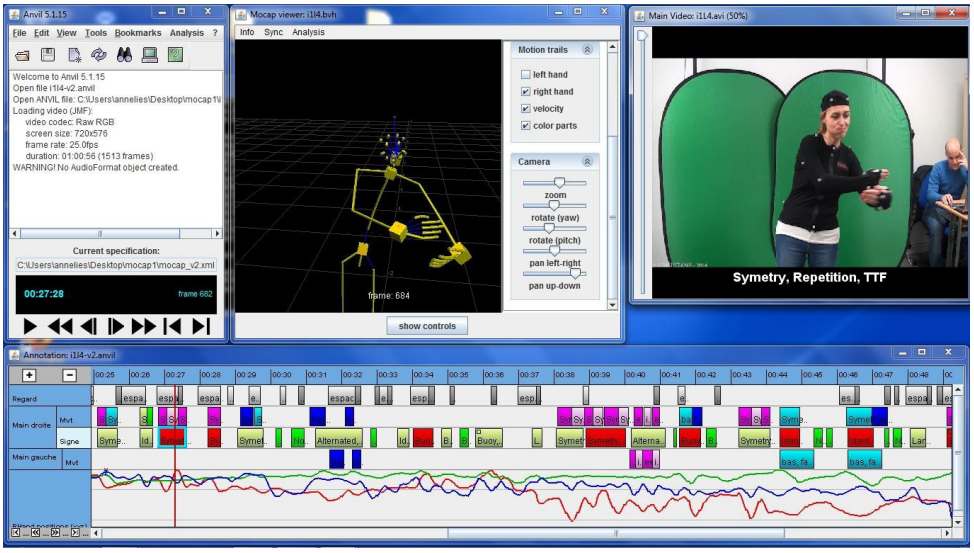


FIGURE 5.3: Manual annotation of MOCAP1 corpus (Benchiheub, 2017) using ANVIL (Kipp, 2001). The annotation tracks entitled in French (bottom) stored information about: eye gaze (*Regard*), manual signs (*Main droite - Signe*) and movements of the two hands (*Main droite/gauche - Mvt*).

5.3 MOCAP1-v2 and the PLmocap library: novel tools for the analysis of Sign Language motion

For now in MOCAP1, the mocap data of only four signers have been made publicly available (referred to as Signers 1, 4, 5 and 6, in Chapters 6, 7, 8, 9 and 10). The mocap data of the two additional signers (Signers 2 and 3) present in this thesis was released as part of a second corpus version: MOCAP1-v2. The motion of the last

two signers (Signers 7 and 8) could finally not be used, because of recording conditions making the data non-recoverable. Beyond the additional data of Signers 2 and 3, further processing of MOCAP1 mocap data has been carried out and is available in MOCAP1-v2. Most of the methods developed in that regard have been publicly released as part of the PLmocap Python library[1]. First, a slightly new representation of the mocap data has been proposed (Section 5.3.1). Then, further processing methods have been implemented in order to normalize the mocap data with respect to structural features including size, shape and posture of the signers (Section 5.3.2). Finally, some visualization tools have been developed for the purposes of SL motion analysis, including the studies presented in Chapters 6, 7, 8, 9 and 10 (Section 5.3.3).

5.3.1 Preprocessing the original mocap data

In the first task of MOCAP1, 25 picture descriptions were recorded. From these 25 mocap recordings, only 24 were taken into account in our version, as one of them was not available for one of the additional signers (Signer 3). Correspondences between the data present in the original version of MOCAP1 corpus and those present in MOCAP1-v2 can be found in Appendix B. Moreover, from the 27 original body markers, we derived 19 secondary markers that optimally describe the major joints of the body, as previously done in many motion analysis studies (Carlson et al., 2020; Troje, 2002a; Toiviainen et al., 2010). This notably eased subsequent visualizations and synthesis computations, using stick figures or Point-Light Displays (PLDs). As shown in Figure 5.4, the derived markers were (L = left, R = right, F = front, B = back): (1) pelvis, (2) stomach, (3) sternum, (4) LB head, (5) LF head, (6) RB head, (7) RF head, (8) L shoulder, (9) L elbow, (10) LB wrist, (11) LF wrist, (12) LB hand, (13) LF hand, (14) R shoulder, (15) R elbow, (16) RB wrist, (17) RF wrist, (18) RB hand, (19) RF hand.

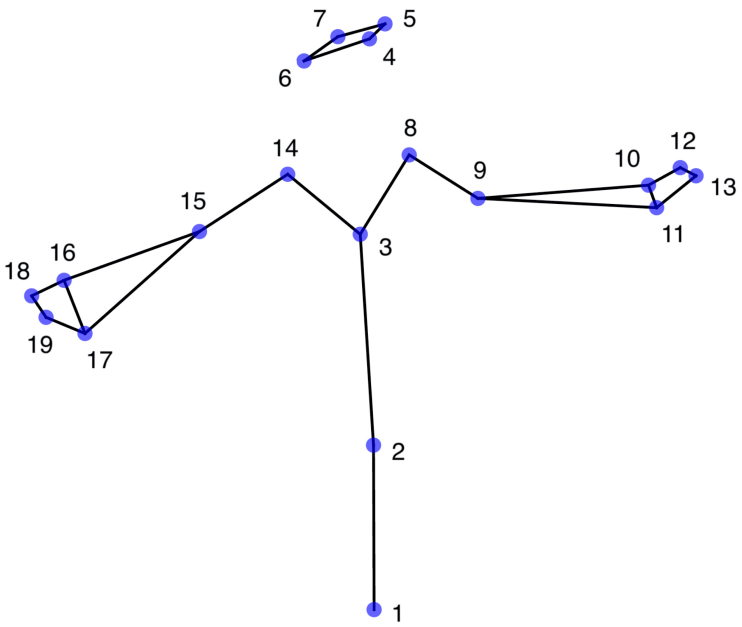


FIGURE 5.4: The 19 markers of MOCAP1-v2 in the "T" reference posture.

[1]PLmocap library. Python code available here: https://github.com/felixbgd/PLmocap.

The 13 original facial markers were not included in this version of the dataset, as our main analyses and processing techniques focused on more global movements of the upper limbs, based on the hypothesis that they may be the most important part of the motion features involved in person identification. Moreover, the social issue of signer identification from virtual signers has been raised (including by professional journals that provide accessible content in SL, such as Media'Pi![2] in France) while animation models rarely, or only partly, convey facial expressions of the signers. Still, further investigations of the role of facial expressions in signer identification, as previously done for intonation (Weast, 2008), could be of great interest.

The positions of the 19 secondary markers were originally provided as 3D Cartesian coordinates in an external reference system (Benchiheub et al., 2016b). Orientations of the body joints were not available from the original corpus. In the dataset presented here, these positions were defined in reference to the pelvis (used as the origin) in order to allow for a better comparison between individuals across mocap examples (see Section 3.2.1 for further discussion about external *versus* body-centered reference systems). PLmocap Python library also provides methods to transform the mocap data from a global coordinate system to a local coordinate sytem (i.e., that defines each marker in reference to his parent marker). For instance, the position of the wrist can be defined in reference to the position of the elbow.

5.3.2 Normalization of structural features

In Section 3.2.1, we discussed how the study of human movements can be influenced by multiple factors, in particular ones related to anthropometric measures. In that regard, two classes of information can be dinstiguished when studying motion science and motion perception: structural and kinematic information. Motion-mediated structural features were defined by Troje et al. (2005) as the invariant information specifying the structure which is put into motion. Structural features thus reveal information about the average posture, and the anthropometric characteristics of the person's body. For structural features to be perceived, PLDs must be in motion. Motion-mediated structural features thus differ from static information, which can be perceived from a static PLD image. However, although they are inferred from moving PLDs, structural features also differ from kinematic ones, which refer to the motion of the body markers itself.

In some of the following studies, the impact of anthropometric measures had to be removed, in order to allow comparing and manipulating the movements of multiple signers *per se*, irrespective of anthropometric differences across individuals (Chapters 6, 7 and 10). Furthermore, in order to evaluate the distinct roles of structural and kinematic information in automatic signer identification (Chapters 8 and 9), the original mocap data (referred to as 'ORI') used in the present study were gradually normalized in three steps (illustrated in Figure 5.5), with respect to size (SI), shape (SH) and posture (POST) of the signers, respectively[3]. Size was defined as the overall length of the body of the signer along the three dimensions. Shape was defined as the individual lengths of the signer's body segments, such as shoulder width, arm length or dimensions of the head. Posture was defined as the average position of the signer's body markers (i.e., how the signer holds his or her body) over all mocap examples. Compared to the two-step normalization procedure proposed

[2] https://media-pi.fr/

[3] The visual perception study of Chapter 8 is an exception: in this study, human observers viewed non-normalized PLDs and the influence of anthropometric cues on their decisions was assessed a posteriori using computational methods.

by Troje et al. (2005), the three normalization steps proposed here allowed us to distinguish the role of postural information from the one of size and shape, which latter are related to the dimensions of one's body, regardless of his or her average posture.

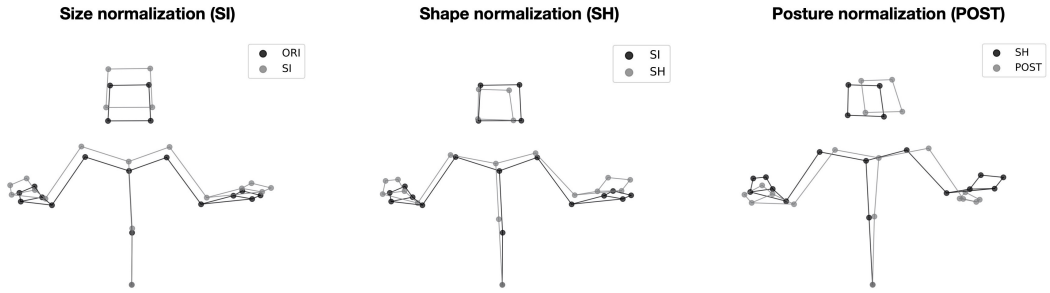


FIGURE 5.5: The three cumulative steps of normalizations of structural features: The stick figures correspond to a given frame of the description of the first picture by Signer 3. For each step, the normalized and non-normalized stick figures are compared. The original motion data (i.e., without any normalization) are referred to as 'ORI'.

Size normalization (SI) was carried out as in Troje et al. (2005). In MOCAP1 corpus, reference "T" postures had been recorded for each signer, at the beginning of each mocap recording (Figure 5.7a). In order to estimate body size differences across signers (see Figure 5.7a), we additionally computed a global reference posture by averaging the reference postures of all signers. The slope of the regression between the 3D positions of each individual reference posture and of the global reference posture was then computed (see Figure 5.6). These slopes defined relative sizes for each signer, which are shown in Table 5.1.

TABLE 5.1: Relative sizes of signers computed from their mocap data, using linear regression.

Signer	Relative size
Signer 1	1.000
Signer 2	1.075
Signer 3	0.924
Signer 4	1.003
Signer 5	0.996
Signer 6	1.003

As shown in Figure 5.7b, after dividing the mocap 3D coordinates of signers by their relative sizes, they all had the same size, while keeping intact shape (i.e., the relative positions of the articulations). Shape normalization (SH) was then computed from the new reference "T" postures of the size-normalized data of each signer (Figure 5.8a). A new global reference "T" posture was defined as the average across signers. Shape-normalized data were obtained by subtracting the individual reference postures from each frame then adding the global reference posture. As shown in Figure 5.8b, after that transformation, all signers had the same reference "T" posture (i.e., same relative positions of the articulations).

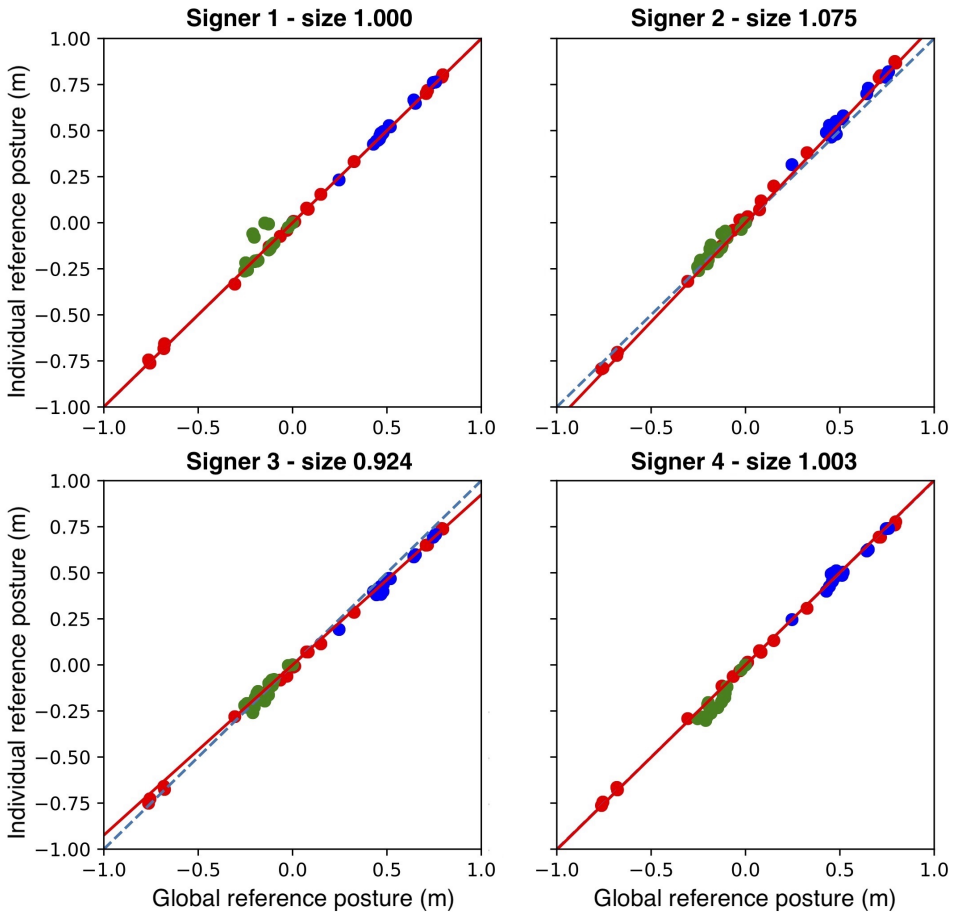
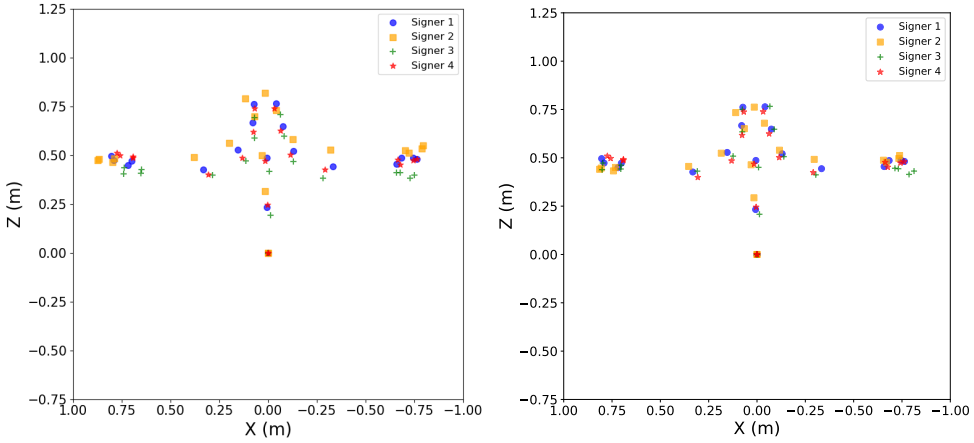


FIGURE 5.6: Scatterplots of the 3D positions of the individual reference posture of each signer, as a function of the global reference posture averaged across all signers. For sake of clarity, computations shown in this figure were made with the first four signers only. Dots represent the marker positions along X (red), Y (green) and Z (blue) axes, respectively. Red lines represent the linear curve estimated by the regression model. The slope of this line defines each signer's relative size. Blue dotted lines represent the linear curve $y = x$. Signers whose regression curve (red) is near the blue dotted line (e.g., Signer 1) have a body size near the global average across signers and thus will not be affected significantly by the normalization.

Finally, posture normalization (POST) was applied to shape-normalized data, which are also size-normalized. Posture-normalized data were obtained by subtracting the average posture of each signer (i.e., averaged over time over all their mocap examples, see Figure 5.9a) from each frame, then adding the global average posture computed over all signers. As shown in Figure 5.9b, after these three normalizations, all signers had the same size, same shape, and same average posture.

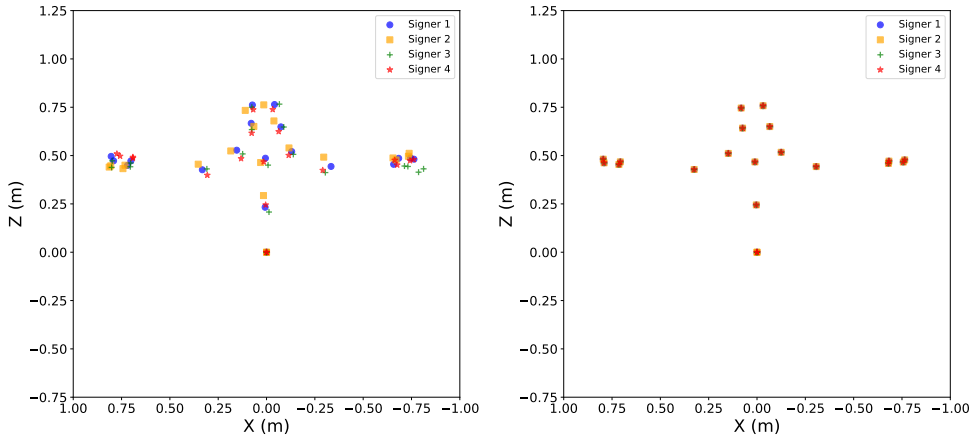
5.3.3 Visualization tools

Various visualization tools have been developed in Python for the aims of the present thesis. They have been publicly released as part of the PLmocap Python library and can be downloaded and re-used here: https://github.com/felixbgd/PLmocap. All methods can be used to visualize motion data from 2D/3D Cartesian coordinates of the body markers. A summary of these functions and of examples of use are shown in Table 5.2 and in Figures 5.10 and 5.11.



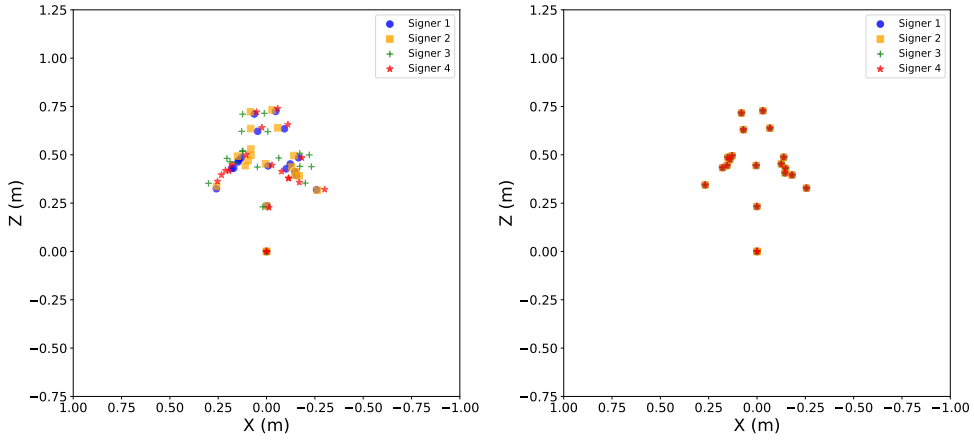
(A) (B)

FIGURE 5.7: Size normalization: original (A) and size-normalized (B) reference "T" postures of the mocap data of the first four signers.



(A) (B)

FIGURE 5.8: Shape normalization: size-normalized (A) and shape-normalized (B) reference "T" postures of the mocap data of the first four signers.



(A) (B)

FIGURE 5.9: Posture normalization: shape-normalized (A) and posture-normalized (B) average postures of the mocap data of the first four signers.

TABLE 5.2: Visualization methods developed as part of the PLmocap library.

Method name	Description	Example of use
`data_viz_3d`	Animated real-time visualization of mocap data in the 3D space, shown as Point-Light Display or stick figure. One specific body marker can be highlighted.	Figures 5.10a and 5.10b
`data_viz_2d`	Animated real-time visualization of mocap data in the 2D frontal plane, shown as Point-Light Display or stick figure. One specific body marker can be highlighted.	Figures 5.10c and 5.10d
`plot_frame`	Visualization of 1 given frame of mocap data in 2D or 3D, shown as Point-Light Display or stick figure.	Figure 5.4
`compare_Nframes`	Comparison of N postures of mocap data in the 2D frontal plane.	Figures 5.7, 5.8 and 5.9
`video_PL`	Generation of a mocap video as Point-Light Display in 2D in either the frontal, sagittal or transverse plane, exported in MPEG-4 format (using `ffmpeg`).	Point-Light videos[4]
`plot_2frames`	Visualization of mocap data in 2D at 2 given frames, in either the frontal, sagittal or transverse plane. The 2 postures are shown as overlapped gray and black stick figures.	Figure 5.11a
`plot_3frames`	Visualization of mocap data in 2D at 3 given frames, in either the frontal, sagittal or transverse plane. The 3 postures are shown side-by-side. Overlapping stick figures of two different individuals can be plotted for comparison.	Figure 5.11b

[4]Mocap video examples used in the present thesis are available here: https://zenodo.org/record/5215804#.YRzcFtMzba4.

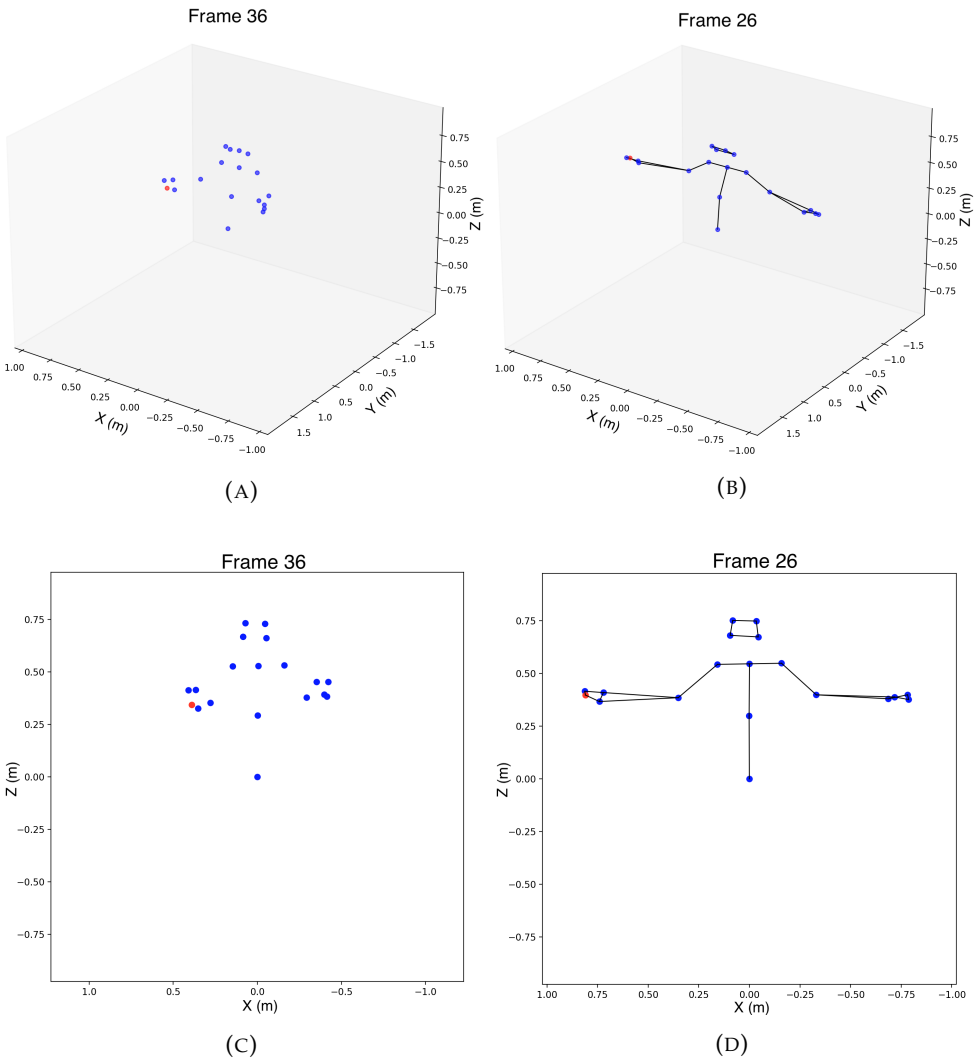
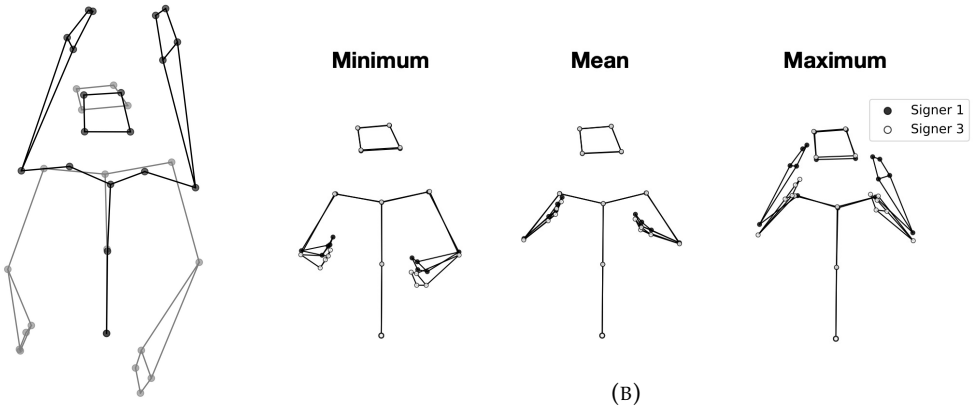


FIGURE 5.10: Real-time visualization of mocap recordings (here shown as screen-shots taken at specific frames during the animation): 3D visualization of the data shown as Point-Light Display (A) or stick figure (B). 2D visualization in the frontal plane of the data shown as Point-Light Display (C) or stick figure (D). In all figures, marker 18 (RB hand) is highlighted.



(A) (B)

FIGURE 5.11: 2-frame (A) and 3-frame (B) visualizations of the mocap data shown as stick figures in the frontal plane. The movement can be described showing key frames (e.g., minimum, mean and maximum of the movement). In the 2-frame example (A), minimum (gray) and maximum (black) of the movement are shown. In the 3-frame example, (B), the movement execution is compared between Signer 1 and Signer 3. These representations were mainly used in Chapter 7.

Chapter 6

How fast is Sign Language? A reevaluation of the kinematic bandwidth

This chapter tackles a traditional problem of signal processing: spectral estimation. As previously mentioned in Section 4.1.1, the goal of spectral estimation is to characterize the frequency content of a temporal signal. In particular, given the hypothesis that a signal is composed of a limited number of frequencies in addition to noise, one main objective of the estimation is to find the location and intensity of these actual frequencies in order to distinguish them from noise. This problem has been widely investigated in various domains, such as neuroscience where techniques have been developed to remove unwanted noise components from neurophysiological recordings (e.g., sensor noise related to electronics, physiological noise related to brain activity other than of interest or environmental noise related to other signals present in the recording area) (De Cheveigné and Simon, 2007; De Cheveigné and Simon, 2008; Narayan, 2018). In the case of mocap data, noise components are mainly due to measurement conditions (e.g., calibration, light interference, marker fixation).

In this thesis, we investigated spontaneous Sign Language (SL) movements using 3D mocap data sampled at high frame rates and thus subject to noise. Filtering the mocap data using a reasonable cutoff frequency could thus provide a meaningful representation of the actual motion for subsequent analyses, while removing unwanted noise components. However, little is known about the kinematic bandwidth of SL, apart from isolated signs. Prior studies examining isolated signs have suggested that SL could be limited to relatively low frequencies. This is unlikely to be appropriate for real-life conditions (i.e., spontaneous productions) where signs are produced faster and are combined with several other rapid motion features (Section 6.1). The study presented in this chapter investigated the spectral content of the MOCAP1 multi-signer mocap corpus of spontaneous signing in French Sign Language (LSF), using Power Spectral Density estimation and residual analysis of the mocap data (Section 6.2). In order to further address the importance of kinematic bandwidth estimation for the purposes of this thesis, bandwidth limited mocap data were used to train a preliminary machine learning model to identify the six signers of the corpus. The performance of the model was assessed, as a function of the used bandwidth (Section 6.3). Results are finally discussed in terms of fundamental findings on the kinematic properties of SL and in terms of application perspectives (Section 6.4).

This chapter is partly reproduced from Bigand et al. (2021b).

6.1 Answers from isolated signs: how incomplete?

Human motion has been shown to lie within a range of low frequencies (i.e., below 20 Hz, see Section 4.1.1). Some studies investigating SL movements have shown that American Sign Language (ASL) isolated signs displayed motion frequencies within even narrower bandwidths than those of non-SL movements, such as 0–6 Hz (Poizner et al., 1986), 0–5 Hz (Sperling et al., 1985) or 0–3 Hz (Foulds, 2004). The limitation of these studies is that SL cannot be restricted to isolated signs taken out of context. Because of coarticulation, the duration of signs is shorter when produced in context rather than isolated (Braffort et al., 2011; Koech, 2006). In addition to lexical signs, SL production is a continuous stream that involves multiple features, including rapid manual (e.g., pointing) and non-manual (e.g., eye gaze) movements. It can therefore be hypothesized that the actual bandwidth of SL motion is wider than previous estimates. As a matter of fact, one of the few studies that have investigated real conversation conditions precisely have indicated that a 5-fps video sampling rate was too low for a comfortable SL conversation (Cherniavsky et al., 2007). Additionally, the studies mentioned above have assessed the signed movements of one individual, which does not account for differences in speed between signers.

The aim of the present study was to overcome these limitations by evaluating the spectral content of spontaneous, continuous, signing and over multiple signers. For that aim, a two-step computational analysis of the mocap data was conducted (Section 6.2). Power Spectral Density estimation and residual analysis were applied to the spontaneous LSF mocap data of the six signers presented in the MOCAP1-v2 corpus (see Section 5.3 for corpus description). To the author's knowledge, the present study is the first to use this computational workflow for SL. Moreover, the high precision of the 3D mocap system used for the recordings allowed for the evaluation of higher frequencies compared with state-of-the-art studies on SL (250 fps *vs.* 30 fps (Foulds, 2004; Sperling et al., 1985)). Note that interesting observations had been made on the frequency content of ASL continuous discourses for avatar animation purposes in McDonald et al. (2016). We further provide a discussion of our results as compared to the latter study in the present thesis (see Section 6.4). Using a different computational approach, we were able to complement the findings of McDonald et al. (2016) with further quantitative analyses of the mocap data in order to determine the bandwidth to be used when modeling SL movements in real-life conditions.

6.2 Frequency content estimation of spontaneous LSF mocap

Similarly to Skogstad et al. (2013), a two-step analysis was conducted in order to define the optimal kinematic bandwidth of SL. First, the frequency content of the movements of each signer was estimated by measuring Power Spectral Density (PSD) (Section 6.2.1). Then, a residual analysis of the mocap data allowed further distinguishing actual motion information from noise components (Section 6.2.2) and thus choosing the optimal cutoff frequency for low-pass filtering (Section 6.2.3). The main findings of these estimations are that SL involves finer motion patterns than non-SL movements, but also faster motion when the SL discourse is produced in a spontaneous, realistic, manner rather than in isolation (Section 6.2.4).

6.2.1 Power Spectral Density estimation

Power Spectral Density (PSD) was estimated using the Welch method (Welch, 1967). Trajectories were split into overlapping segments over time, then the periodogram (i.e., magnitude squared of the windowed Discrete Fourier Transform) of each segment was computed. The PSD estimates were finally obtained by averaging the periodogram values over all segments. The present analysis was carried out using a Hann window of size 250 (1 sec), with 66% overlap.

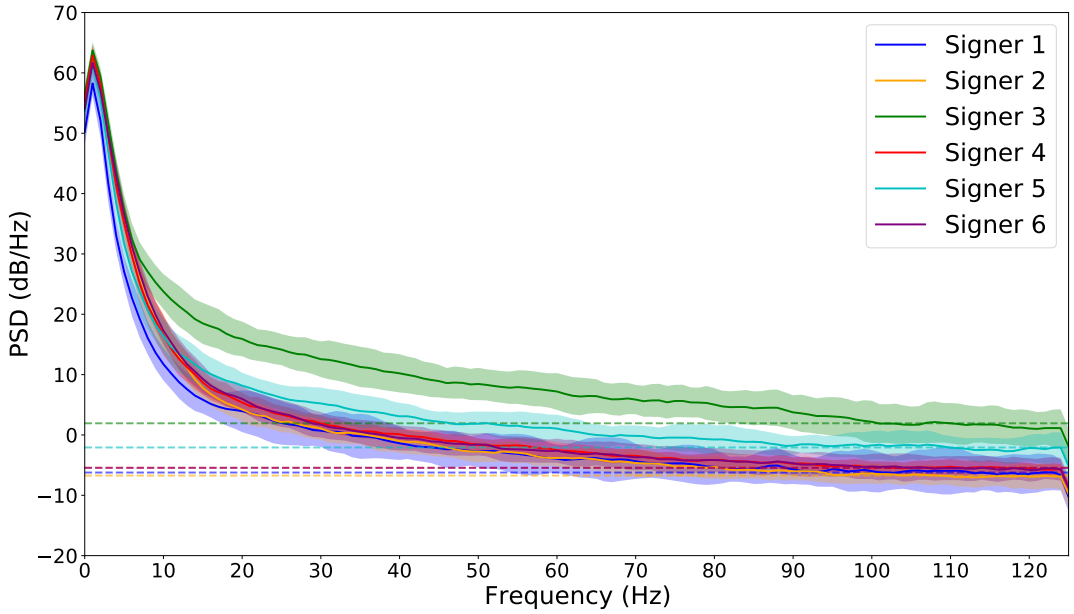


FIGURE 6.1: Power Spectral Density estimation of the mocap data of the six signers. Dashed horizontal lines indicate the noise floor estimate, for each signer. Error shaded regions indicate standard deviations over mocap examples.

The PSD estimates of the mocap data of the six signers are shown in Figure 6.1. The PSD values were averaged over body markers and over mocap examples. Noise floors were estimated by a visual analysis of the PSDs. Signers 1, 2, 4 and 6 reported similar noise floor (-6 dB/Hz), while it was higher for Signer 5 (-2 dB/Hz) and additionally higher for Signer 3 (+2 dB/Hz). Most of the power distribution lies between 0 and 5 Hz, with a 3-Hz peak, for all signers. Still, higher frequencies seem to contribute significantly as the associated PSD values are distinct from the noise floor up to 50 Hz (or higher, e.g., for Signer 3).

6.2.2 Residual analysis

In order to further understand whether these higher frequencies related to actual motion information or to measurement noise, a residual analysis was conducted in the same way as in Winter (2009). This method consists of measuring the average difference between the unfiltered and filtered signal, over several cutoff frequencies. In this study, the mocap data were low-pass filtered using a fourth-order Butterworth filter.

Results of the residual analysis between unfiltered and filtered mocap data are displayed in Figure 6.2. As for PSD estimation, the residual values were averaged over markers and examples. The estimates of noise residuals were obtained by defining the regression line from 40 Hz to $fps/2$ Hz. Indeed, theoretically, the residual

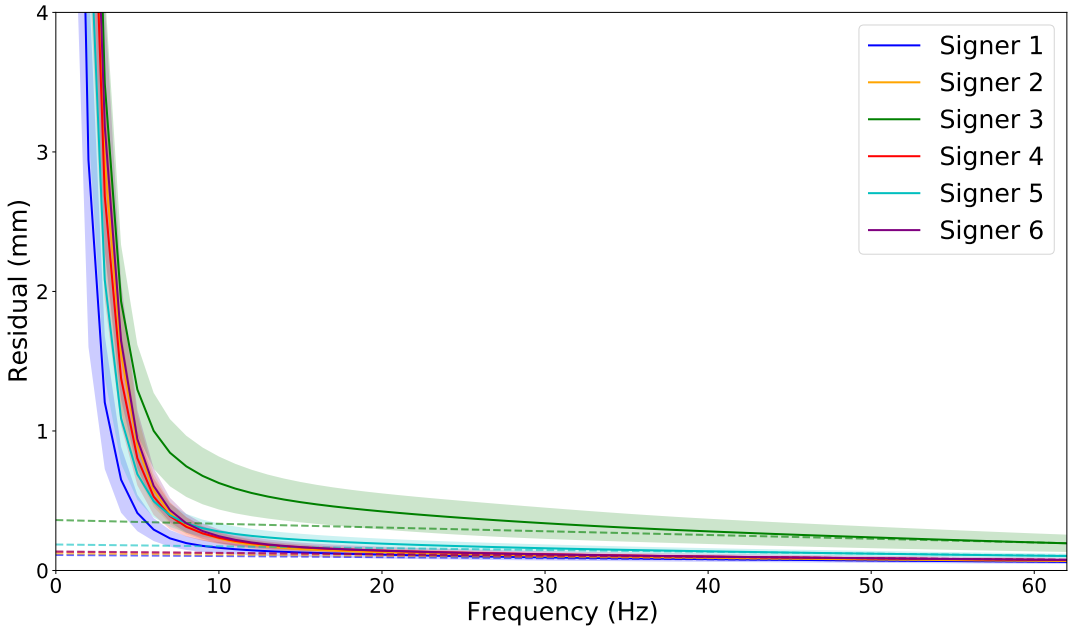


FIGURE 6.2: Residual plot between the unfiltered and filtered mocap data of the six signers, as a function of the filter cutoff frequency. Dashed lines indicate the noise residual estimate, for each signer. Error shaded regions indicate standard deviations over mocap examples.

curve of noise is a linear curve from an intercept at 0 Hz to the ($fps/2$ Hz, 0 mm) point.

This 0-Hz intercept provides an estimate for the Root Mean Square (RMS) of noise, following the definition of Winter (2009). In other words, this value reflects the mean displacement of sensors caused by measurement noise. Estimated RMS of noise were similar for Signer 1 (0.11 mm), Signer 2 (0.12 mm), Signer 4 (0.14 mm) and Signer 6 (0.14 mm). Higher values were reported for Signer 5 (0.19 mm) and additionally higher for Signer 3 (0.37 mm). These results confirmed the PSD estimation, suggesting that the mocap data of Signers 3 and 5 were the noisiest.

Interestingly, Figure 6.2 shows that the residual values relating to Signer 5 (cyan curve) are lower than those of Signers 2, 4 and 6 (yellow, red and magenta curves, respectively) in low frequencies (below 8 Hz), but higher in high frequencies (above 8 Hz). The latter high-frequency comparison is in line with the noise RMS calculations. This suggests that most of the actual motion information of Signer 5 may be in lower frequencies (i.e., slower movements), while his mocap recording is noisier.

6.2.3 Choosing an optimal bandwidth

We then assessed different cutoff frequencies in order to define the optimal kinematic bandwidth of our data. Based on prior work, three cutoff frequencies were compared: 6, 12 and 25 Hz. The lower frequency relates to a 1-mm residual[1], 1-mm deviations being imperceptible for arbitrary hand motion (Skogstad et al., 2013). The upper one is the frequency for which the residual equals the noise RMS[1]. Using this cutoff value, the signal distortion should equal the amount of noise allowed through (Winter, 2009). Finally, 12 Hz is an intermediate value of great interest as it would be the highest cutoff frequency possible with most video systems (sampled at 24 fps), following the Nyquist-Shannon theorem (Nyquist, 1928; Shannon, 1949).

[1]When different frequencies were possible across individuals, the maximum frequency was chosen, in order to minimize signal distortion.

When looking at slow motion (Figure 6.3), the 6-Hz and 12-Hz filters seem to be optimal solutions for denoising the data. The 6-Hz filter might even be slightly more promising, as it filters out more artifacts than the 12-Hz one. However, this finding is not confirmed with rapid motion (Figure 6.4), where filtering at 6 Hz cancels important fast movements. For instance, the residual values relating to Signer 3 almost double from 6 Hz (0.56 mm) to 12 Hz (1.00 mm). Interestingly, the 12-Hz filter smoothes out most of the signal, while keeping fast oscillations intact. Filtering at 25 Hz instead of 12 Hz does not seem to add substantial information and differences in the residuals are negligible (M = 0.08 mm, SD = 0.04).

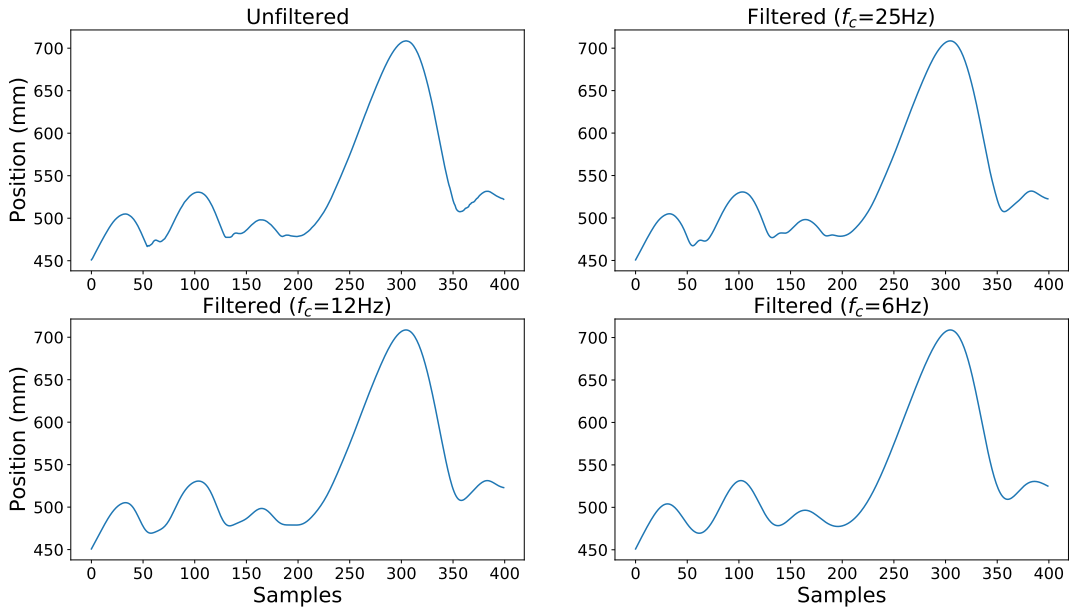


FIGURE 6.3: Example of slow motion: Z-axis trajectory of the right hand of Signer 5, for mocap example 5. Subplots allow for comparison between unfiltered and filtered (f_c = 25, 12 or 6 Hz) mocap data.

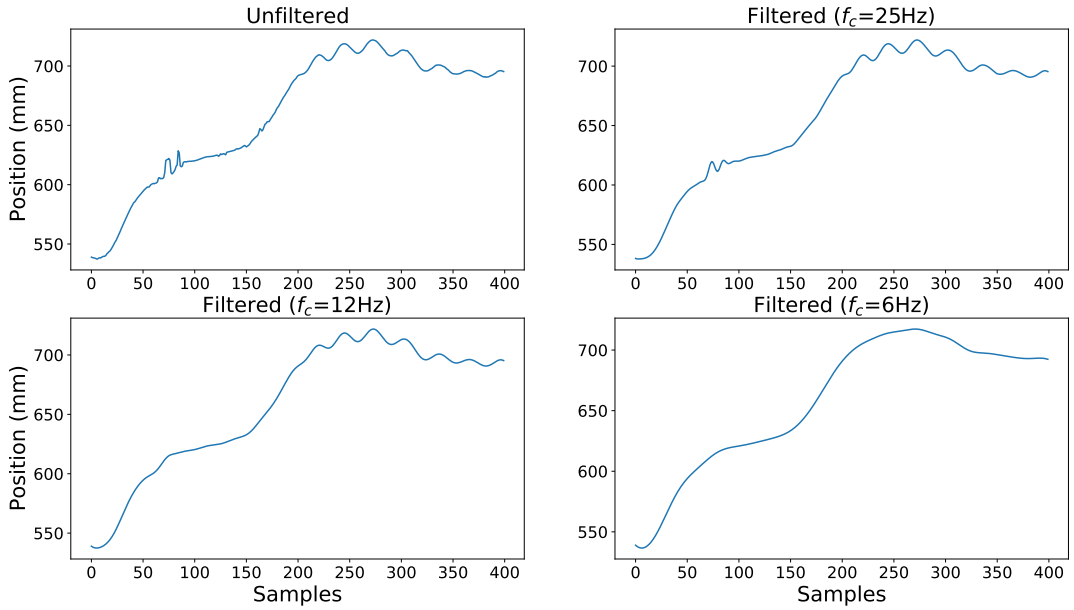


FIGURE 6.4: Example of rapid motion: Z-axis trajectory of the right hand of Signer 3, for mocap example 17. Subplots allow for comparison between unfiltered and filtered (f_c = 25, 12 or 6 Hz) mocap data.

According to these results, the conclusions about the three proposed bandwidths for SL motion are summarized as follows:

- 0–6 Hz: The main movements are captured and the noise artifacts are drastically reduced. However, rapid motion is filtered out, which distorts the motion representation for faster signers.
- 0–12 Hz: The fastest movements are captured while the noise artifacts are still importantly reduced.
- 0–25 Hz: Noise artifacts are allowed through, while the additional information compared with a 0–12-Hz range is negligible, as regards residual values.

6.2.4 Spontaneous Sign Language movements: finer and faster

Based on the combined results of residual analysis and data visualization, 0–12 Hz was found to be a reasonable bandwidth for our SL mocap dataset. This is noticeably wider than the 0–3-Hz (Foulds, 2004) or 0–6-Hz (Poizner et al., 1986) previously reported bandwidths. More specifically, a 0–6-Hz range was not able to account for the rapid signed motion. This range was associated with a 1-mm residual, which was reported to be an unsignificant distortion for rapid arbitrary motion (Skogstad et al., 2013). A 1-mm deviation may thus not be negligible for SL motion, suggesting that this latter contains finer movements. This is consistent as, compared to arbitrary hand motion, SL obeys to specific linguistic constraints and requires precise movements of the hands and fingers for comprehensibility (Poizner et al., 1981). The precision of the analyzed motion may also have been caused by the high level of expertise of the six signers, all being highly fluent in LSF.

It was not clear whether a 0–25-Hz bandwidth actually provided additional information about the real motion rather than noise. Although it was negligible here, it might be possible that higher frequencies relate to actual motion, particularly for fast signers. Further work measuring eye movements could also refine the relevance of a wider bandwidth.

6.3 Why care about kinematic bandwidth estimation?

We further assessed the importance of kinematic bandwidth estimation for the purposes of the present thesis. First, the estimation can impact the extraction of high-level features derived from the mocap data, such as velocity and acceleration (Section 6.3.1). Moreover, as such features are often the basis of machine learning models of motion, wrong estimations of the kinematic bandwidth can further cause misleading results of the model (Section 6.3.2). These observations emphasize the need for defining an optimal kinematic bandwidth when developing analyses and machine learning of SL motion, in particular for the main focus of this thesis: signer identification (Section 6.3.3).

6.3.1 The effect of kinematic bandwidth estimation on feature extraction: the example of velocity and acceleration

Computational models rely on features derived from the trajectories of markers, such as velocity or acceleration, which are highly sensitive to prior filtering of the mocap data. As shown in Figure 6.5, the amount of noise is amplified at each step

of differentiation. More interestingly, without filtering, the acceleration data are almost not readable, which may cause wrong interpretations of inter-individual differences (e.g., acceleration peak of Signer 5, instead of Signer 3). Person identification is particularly suited to further illustrate this issue, as wrong interpretations of inter-individual differences may cause wrong predictions of the identified person. Moreover, person identification from motion recently raised important social issues about the confidentiality of signers in SL (see Section 1.3). Therefore, using signer identification as an example, a machine learning model was further trained and its performance was assessed, as a function of the used bandwidth.

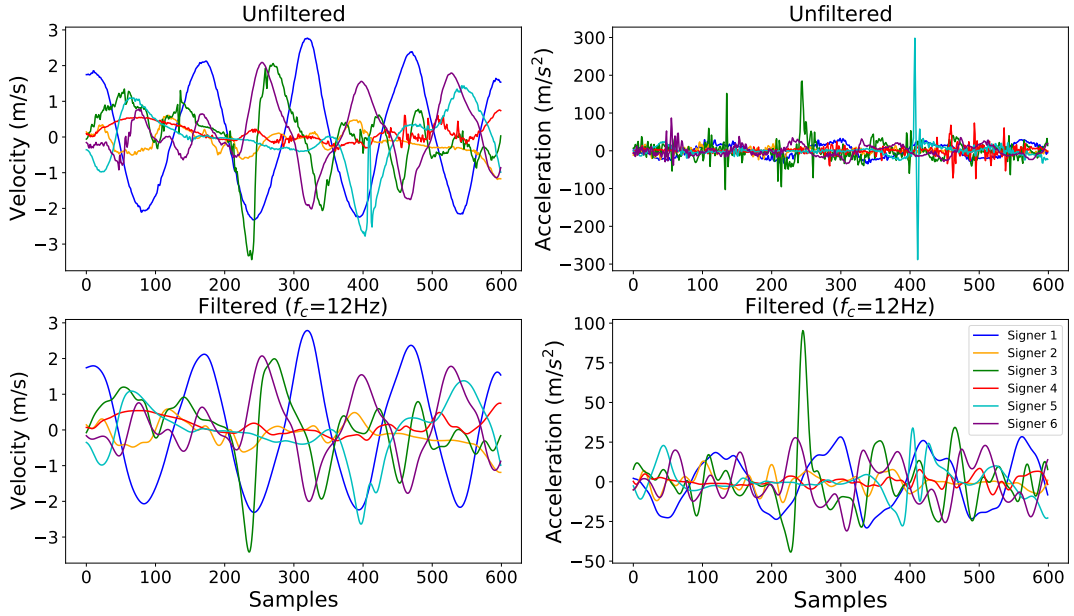


FIGURE 6.5: Z-axis velocity (left) and acceleration (right) curves of the right hand, for all signers, for mocap example 1. Subplots allow for comparison between unfiltered and 12-Hz filtered mocap data. Note the scale difference of the Y dimension for acceleration, which reflects the unrealistic values caused by unfiltered data.

6.3.2 Further implications for machine learning models of motion

The model was designed using a statistical-based approach. Statistics (mean and standard deviation) were measured from temporal features of each mocap example. The used temporal features were a combination of local position, velocity and acceleration. The identification step consisted of applying Principal Component Analysis to the statistics ($\mu_{pos}, \sigma_{pos}, \mu_{vel}, \sigma_{vel}, \mu_{acc}, \sigma_{acc}$) of all examples and finally training a multinomial logistic regression model on the extracted Principal Components (PCs). The model was trained iteratively on N-1 (23) examples for each signer, and the remaining 1 observation was used as test exemplar. Performance was computed as an average over the 24 test iterations. Note that this model was preliminary and was used to shed light on the potential effects of kinematic bandwidth estimation on our subsequent investigations. The further development of machine learning models for automatic signer identification will be presented in Chapter 9.

In order to illustrate the impact of the bandwidth estimation on the model, the first extracted PC was analyzed. This PC was highly correlated with global dynamic statistics ($\sigma_{vel}, \sigma_{acc}$) for both unfiltered ($r(16416) = .63, p < .001$) and 12-Hz filtered

($r(16416) = .65, p < .001$) inputs[2]. Yet, when trained on this PC, the model provided different results between the two filtering conditions. Figure 6.6 displays the projections of all the mocap exampels onto the first two PCs and the model confusions when trained on PC1.

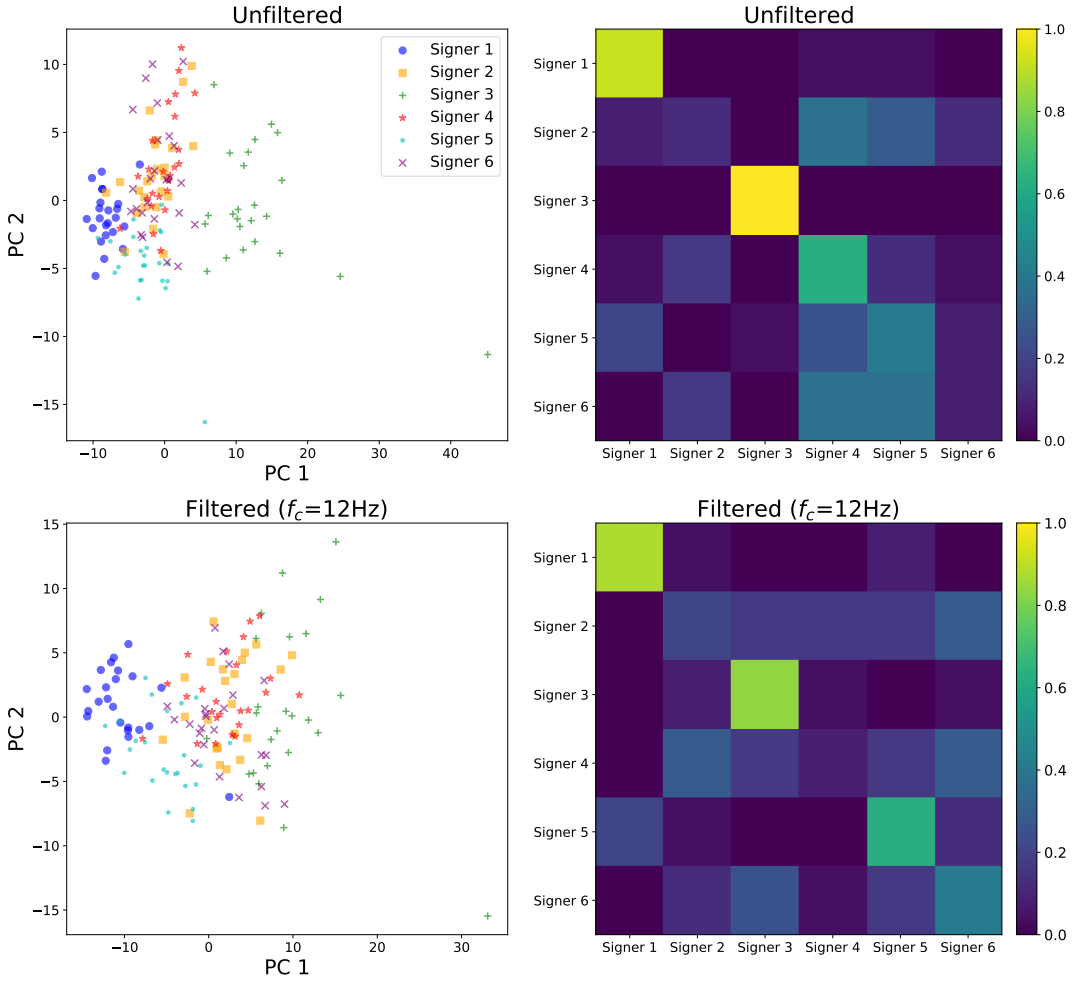


FIGURE 6.6: Left column: projection of all the mocap examples over the first two PCs extracted by the model. Right column: confusion matrix showing the identifications of the model (averaged over 24 tests), when trained only on PC1. The two rows allow for comparison between unfiltered and 12-Hz filtered mocap data.

For unfiltered mocap data, the PC1 weights of Signer 5 (M = -0.15) were confused with those of Signer 2 (M = -1.43), Signer 4 (mean = -2.72) and Signer 6 (M = -1.04). By contrast, Signer 1 had lower weights (M = -7.72) and Signer 3 was surprisingly highly separated from the 5 other signers (M = 13.07). The performance of the model trained on PC1 confirmed this idea, as Signer 1 (91.7 %) and Signer 3 (100%) were correctly identified, by contrast with Signer 5 (41.7 %). However, when applying a 12-Hz filtering, the PC1 weights of the mocap data were different. Signer 1 still had the lowest weights (M = -10.4), but signer 5 (M = -5.05) was separated from Signer 2 (M= 2.17), Signer 4 (M = 1.99) and Signer 6 (M = 1.67). The highest weights reported were still those of Signer 3 (M = 9.62), but the gap with other signers was lower. This time, the model succeeded in identifying Signer 5 with a higher accuracy (62.5%) based

[2]By contrast, the correlation between this PC and static statistics (μ_{pos}) was not significant for unfiltered ($r(8208) = -.02, p = .14$) and 12-Hz filtered ($r(8208) = -.02, p = .18$) inputs.

on his dynamic differences, and Signer 3 was still identified (83.7%) but potentially based on a more realistic interpretation of PC1. Finally, when the model was trained on 6-Hz filtered mocap data, the identification accuracy significantly decreased for the fastest signer, Signer 3 (70.8 %).

6.3.3 A crucial first step in investigating signer identification

These results reflect the observations made in Section 6.2. With the wrong bandwidth estimation, our model may have misidentified Signer 5 because of slower but also noisier motion data, compared to Signers 2, 4 and 6. Similarly, the separation of Signer 3 from other signers was surprisingly wide, which may have been caused by the fact that the mocap of Signer 3 contained the fastest but also noisiest data. The results using a 6-Hz filter also confirm that a 0–6-Hz range distorts rapid motion and thus provides an incomplete representation for models. Based on the example of PC1, 12-Hz filtered data provided the best representation to correctly differentiate between the dynamics of the six signers.

6.4 Conclusion and discussion

The present study provides a new estimate of the kinematic bandwidth of Sign Language using computational methods, as was done for gait and hand motion (Skogstad et al., 2013; Winter, 2009). Compared to prior work on SL isolated signs or fingerspelling (Poizner et al., 1986; Foulds, 2004; Sperling et al., 1985), our results suggest that SL motion contains higher frequencies (0–12 Hz). In the present study, signers had freely described pictures in French Sign Language without any constraints in time, signs or structure. These results thus support the hypothesis that signing may be faster when it is done in context, rather than when it is isolated (Braffort et al., 2011). Furthermore, the use of the mocap data of six signers allowed for more generalization, compared to prior work. Interestingly, our estimate is in line with the prior work of McDonald et al. (2016) on continuous SL, which supported 0–12-Hz as a proper kinematic bandwidth for animating virtual signers with realistic movements. More precisely, the latter study distinguished two bandwidths. Although only the movements within a 0–4-Hz spectral range may convey linguistic meaning, movements within a 4–12-Hz range may still be related to actual human motion rather than noise and should thus be taken into account for enlivening virtual signers. In our study, we focused on a different purpose than animation (e.g., machine learning classification task) but our results yielded similar conclusions. Indeed, the performance of our machine learning model was optimal when using a 0–12-Hz bandwidth, which further emphasizes the need for a correct filtering of mocap data when designing SL models. For technological application purposes, these results support the potential for SL motion data extracted from videos at 24 fps (Cao et al., 2019). Despite the fact that estimating movements from videos remains limited to two dimensions, it may properly capture spectral information, following the Nyquist-Shannon theorem (Nyquist, 1928; Shannon, 1949). These outcomes call for additional research further investigating the kinematic bandwidth of SL across other signers and within different linguistic contexts.

Chapter 7

Decomposing Sign Language into Principal Movements

The high number of body segments and the variety of linguistic structures involved in SL movements make their modeling particularly challenging. Compared to other human actions with simpler temporal structures, such as walking, skiing or juggling, the extent to which SL complex movements could be reduced to a lower dimensional space remains unclear (Section 7.1). This chapter tackles this problem by testing Principal Movement (PM) decomposition (see Section 3.2.4) on full-body 3D mocap data of spontaneous SL. For that aim, Principal Component Analysis (PCA) was applied to the mocap data described in Chapter 5 in order to determine the PMs of French Sign Language (LSF) (Section 7.2). We aimed to provide qualitative descriptions of these PMs, and to question whether the extracted PMs were signer-specific or shared by all signers (Section 7.3). The main findings of this study are then discussed as compared to prior work on non-SL movements and on SL movements made in isolation (Section 7.4).

This chapter is partly reproduced from Bigand et al. (2021a).

7.1 Fundamental and application perspectives from Principal Movement decomposition

Investigating new computational models for the description of SL movements presents a two-fold interest. First, it would allow gaining fundamental insights into the structure of SL movements. Second, it could be used in order to improve technological tools dedicated to SL. As previously mentioned (see Chapter 4), the progress of mocap systems has enabled researchers to study human movements, based on large amounts of data from multiple body markers of moving persons. Still, most studies have relied on specific variables of the movements defined by researchers in order to anlyze human movements, in particular in SL. For instance, the mocap analyses of Malaia et al. (2008) and Malaia and Wilbur (2012) have assessed the relationship between pre-selected kinematic features (e.g., peak speed or acceleration) and linguistic features of American and Croatian SL. Similarly, the motion features conveying prosodic variation of poetic sequences in Catteau et al. (2016) have been unveiled by comparing the behavior of candidate kinematic features (i.e., thought to have a role in the prosodic variations) and the poetic annotations.

By contrast, some studies investigating non-SL motion have taken another approach to describe the movements from a holistic perspective, using PCA. Unlike the analysis of pre-selected variables, this data-driven method has allowed for disentangling how complex multi-segmental movements, such as gait, karate, skiing or juggling, were structured, without any a priori hypotheses (see Section 3.2.4). To the

author's knowledge, such a holistic evaluation of the full-body movements of spontaneous SL has not been proposed yet. The only insights gained into potential PM decompositions of SL movements have been obtained from gestures of the dominant hand of signers during the production of highly constrained isolated ASL signs (Yan et al., 2020). In the latter study, 11 PMs accounted for 95% of variance in the original hand movements produced when signing isolated alphabetical letters or numbers from 0 to 10. Producing ASL letters or numbers with the hand is only a reduced part of particular ASL signs. Moreover, as pointed out repeatedly in this thesis, SL productions made in isolation hardly provide complete descriptions of how SL is used by signers in real-life conditions, even for more complex signs than letters or numbers. We precisely demonstrated that spontaneous SL mocap recordings could reveal faster movements than prior estimates made on isolated signs (see Chapter 6). Finally, SL motion involves far more body parts than just the dominant hand, including the other hand, but also torso, head, shoulders and arms. Interestingly, the mocap data used in the present thesis include full-body trajectories compared to Yan et al. (2020), which have investigated the kinematics of the dominant hand only. Inversely, Yan et al. (2020) have applied PCA to precise recordings of finger gestures, while our mocap data were recorded on various upper limbs but only included global motion data of the wrists and hands. Applying PM decomposition to the full-body mocap data provided by MOCAP1 (Chapter 5) would thus interestingly complement these prior analyses made on a limited subset of ASL signs.

PM decomposition is a data-driven method that may provide unexpected fundamental insights into how complex movements of SL are structured and how computational models could automatically decompose SL into simpler, elementary, movements. Moreover, this method is also of particular interest for application purposes because it allows for dimensionality reduction. Computational models of SL motion could rely on dense mocap datasets by processing only a reduced subset of PMs while keeping most of the information about the original movements.

7.2 Methods

Similarly to Chapter 6, this study was conducted on the LSF mocap data of the six signers of MOCAP1-v2 (Section 7.2.1). In order to assess the extent to which PMs of SL may be specific to each signer or shared by all signers, common PMs were extracted from the mocap data of all signers, while individual PMs were extracted from the mocap data of the six signers separately (Section 7.2.2).

7.2.1 Mocap data processing

The mocap data were low-pass filtered using a 4th-order Butterworth Filter with a cutoff frequency of 12 Hz, following the estimations of Sign Language kinematic bandwidth presented in Chapter 6. From each of the 24 original recordings provided by MOCAP1-v2 (see Chapter 5), one mocap recording unit with the duration of 5 seconds was extracted from the beginning of the utterance, irrespective of the semantic content. Each mocap recording unit was thus related to a different picture description in LSF. This resulted in 24 mocap examples per signer, of 5-second duration each (see examples in Videos 7.1 to 7.6).

The movements of each individual signer (i.e., the concatenation of their 24 mocap examples) were described in a matrix containing 30,000 posture vectors (rows) defined by the 3D coordinates of the 19 markers (columns) at each time frame t:

$$\boldsymbol{p}(t) = [x_1(t), y_1(t), z_1(t), ..., x_{19}(t), y_{19}(t), z_{19}(t)] \tag{7.1}$$

where $\boldsymbol{p}$ is the posture vector.

Based on prior definitions of a "posture space" (Troje, 2002a; Federolf et al., 2014), movements of the signers were thus described as time series of postures, in a 57-dimensional space. Postures were normalized in order to filter out anthropometric differences. Each signer's average posture was subtracted from posture vectors at each frame, and replaced by the average posture over all signers, similarly to Troje et al. (2005). The six distinct matrices of individual signers were used to analyze the movements of each signer separately. In order to test the extraction of common motion patterns across individuals, these matrices were also concatenated into a $(180,000 \times 57)$ matrix containing the posture vectors of the six signers.

7.2.2 Principal movements

As outlined by Troje (2002a), a set of human posture vectors can be highly redundant, because of biomechanical constraints and motor control laws. PCA is a mathematical method that appears well suited to measure this redundancy. PCA decomposes the original matrix into a set of uncorrelated Principal Components (PCs), which are the eigenvectors of the covariance matrix of the original data (Abdi and Williams, 2010). This new PC space is organized so that the first PCs maximize the amount of the variance in the data, which makes it possible to conduct analyses on a reduced set of PCs. Here, we applied PCA to the centered posture matrix (i.e., we subtracted the mean from the matrix columns). Based on the 57×57 covariance matrix of the posture data, PCs (or eigenvectors) and their respective eigenvalues were computed. The normalized egeinvalues indicated the percentage of variance explained by the related PCs. We were then able to define each posture $\boldsymbol{p}(t)$ as a linear combination of the PCs:

$$\boldsymbol{p}(t) = \boldsymbol{p}_0 + \sum_{i=1}^{K} w_i(t)\boldsymbol{V_i} \tag{7.2}$$

where $w_i(t)$ is the projection of the posture vector $\boldsymbol{p}(t)$ onto the i^{th} PC (or eigenvector) $\boldsymbol{V_i}$. $\boldsymbol{p}_0$ is the average posture vector over time. K is the number of PCs used to reconstruct $\boldsymbol{p}(t)$, $K \in [1, 57]$.

The PCs captured directions of high variability in the original movements. In other words, PCA allowed decomposing the movements of multiple inter-correlated markers in three dimensions into elementary one-directional movements, called principal movements (PMs). First, a PCA was applied to the whole dataset containing the movements of all signers. This first PCA enabled us to extract common PMs, shared by the six different signers. Second, a PCA was applied to the mocap data of each signer, separately. This second PCA provided individual PMs specific to each signer. The number of PMs needed to account for most of the variance in the SL movements was assessed, and compared between individual and common PMs. Moreover, by projecting separately each specific PC back into the original 3D space, PMs were resynthesized using stick figures (equation 7.3). This allowed characterizing the PMs and comparing the motion patterns they described over signers.

$$\boldsymbol{p}_{v_i}(t) = \boldsymbol{p}_0 + w_i(t)\boldsymbol{V_i} \tag{7.3}$$

where $\boldsymbol{p}_{v_i}(t)$ is the posture vector that describes the movements of the i^{th} PM. $w_i(t)$ is the projection of the posture vector $\boldsymbol{p}(t)$ onto the i^{th} PC (or eigenvector) $\boldsymbol{V_i}$.

7.3 Results

First, common PMs of the spontaneous LSF movements provided a low-dimensional space explaining most of the variance in the original movements (Section 7.3.1). Then, differences in the execution of common PMs were found across signers (Section 7.3.2). Finally, individual PMs were compared with common PMs in terms of the motion patterns they quantified, and in terms of the number of PMs needed to explain a sufficient amount of variance in the original movements (Section 7.3.3).

7.3.1 Common Principal Movements

As shown in Figure 7.1, the first eight common PMs explained most of the variance in the mocap data containing the 24 examples of the six signers. Combined, they explained 94.9% of the cumulative variance.

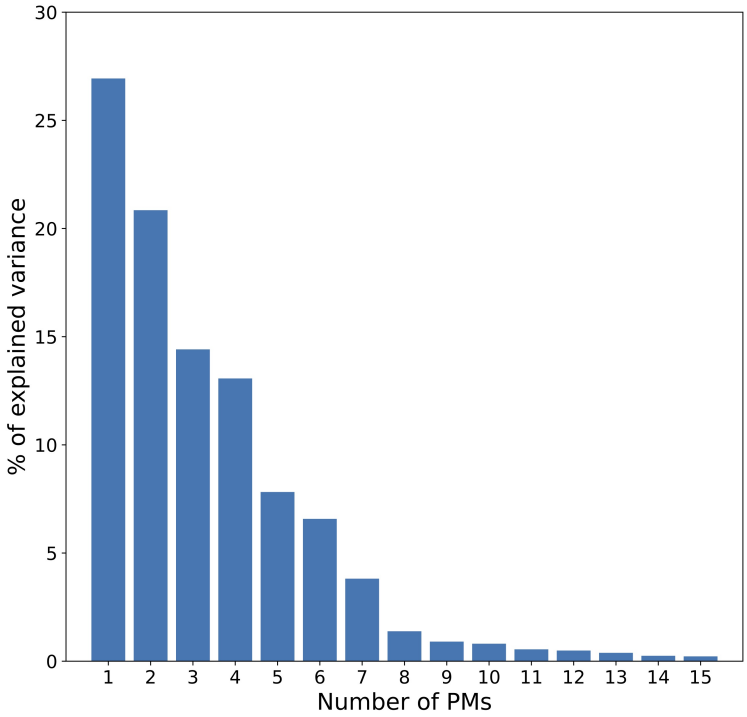


FIGURE 7.1: Variance explained by the first 15 common PMs.

The first eight common PMs (Videos 7.7 to 7.14) are shown in Figure 7.2 and are described in details in Table 7.1. In summary, the first eight PMs were mainly defined as motion patterns visible in the frontal and sagittal planes. PM1 to PM4 quantified movements of the two hands along the vertical, anteroposterior and mediolateral axes, as well as upper-body rotation around the vertical axis. PM5 and PM6 quantified movements of the hands, similarly to the first three PMs, but with parallel rather than opposite movements of the hands, or vice versa. For instance, parallel mediolateral movements of the hands were found in PM5, jointly with lateral sway, compared with the opposite movements of PM3. Similarly, opposite vertical movements of the hands were found in PM6, compared with the parallel movements of PM1. Higher-order PMs (PM7 and PM8) extracted finer movements, such as flexion

of the wrists or shoulder abduction. Moreover, PM8 reported covarying movements between the head and the arms of the signers. For instance, a low negative weighting of PM8 was related to high flexion of the head and high abduction of both shoulders.

TABLE 7.1: Characterization of the first eight common PMs. EV is the Explained Variance in original movements.

PM	EV (%)	Description
1	26.9	Vertical parallel movement of the hands.
2	20.9	Anteroposterior parallel movement of the hands.
3	14.4	Mediolateral opposite movement of the hands achieved with internal rotation of the arms.
4	13.1	Upper-body rotation around the vertical axis, jointly with internal rotation of the arms.
5	7.8	Parallel shift of the two hands along the mediolateral axis, jointly with slight mediolateral sway of the upper-body.
6	6.6	Opposite vertical movement of the two hands.
7	3.8	Upper-body rotation around the vertical axis, jointly with wrists flexion.
8	1.4	Abduction of both shoulders while elbows are flexed, jointly with head tilt.

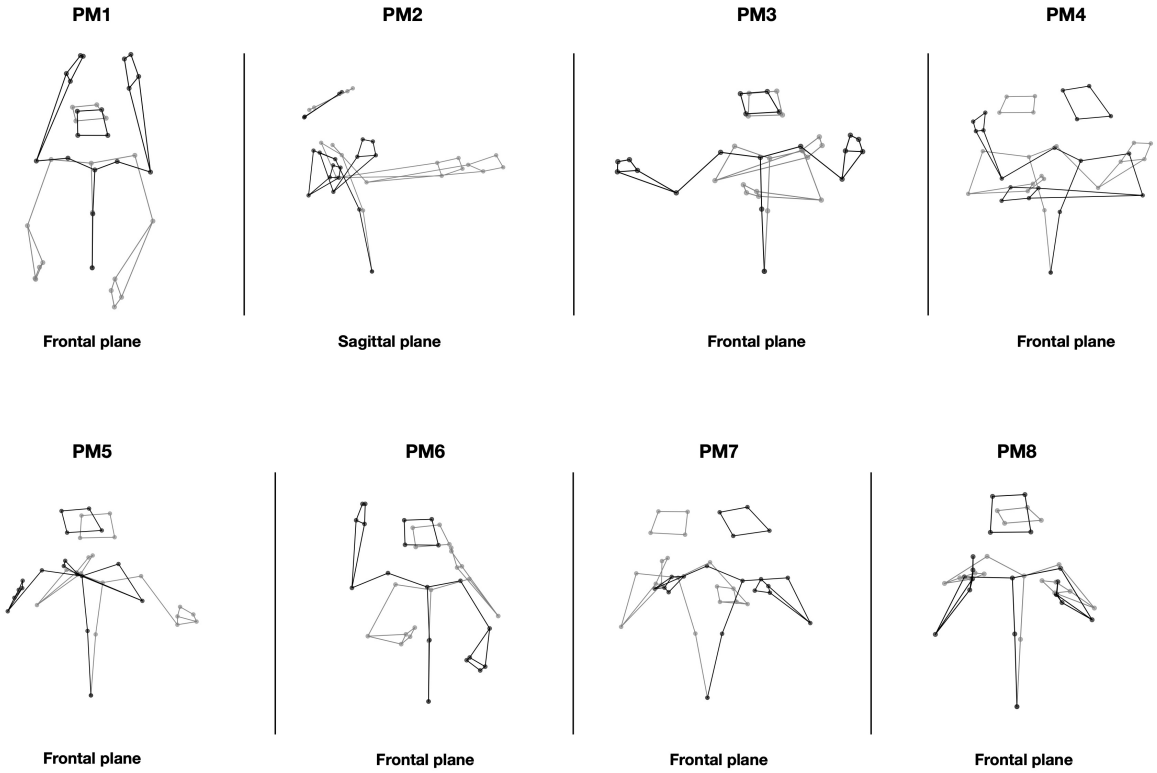


FIGURE 7.2: The first eight common PMs. Stick figures represent the PM at the time instants corresponding to the minimum (gray) and the maximum (black) PM weighting, across signers and examples. PMs are displayed in their main plane of motion (e.g., frontal or sagittal). For sake of visibility, the PM weightings of PM2 were attenuated with a factor 0.75.

7.3.2 Inter-individual differences in the execution of common Principal Movements

Common PMs were used differently across signers. To illustrate that, the execution of the first four PMs was compared between Signer 1 and Signer 3, as shown in Figure 7.3. For instance, in PM1, Signer 1 reported higher positive weightings (of maximum value 0.50) while Signer 3 reported lower negative ones (of minimum value -0.60). This corresponds to higher position of the two hands for Signer 1 and lower position of the two hands for Signer 3, respectively. In PM2, Signer 1 showed mainly positive weightings (M=0.11, SD=0.24) while Signer 3 showed mainly negative ones (M=-0.14, SD=0.17), revealing more remote position of the hands from the body, in average, for Signer 3 than for Signer 1. PM3 reported another pattern of differences, as the two signers presented similar mean and minimum postures, but a consequent gap between their maximum postures. At the lowest level, hands of the signers were similarly close, while at the highest level, the hands of Signer 3 were spaced wider apart, compared with Signer 1. In PM4, Signer 1 showed mainly positive weightings (M=0.03, SD=0.09) while Signer 3 showed mainly negative ones (M=-0.11, SD=0.08), revealing slightly higher leftward rotation of the body for Signer 1 and rightward rotation for Signer 3, respectively.

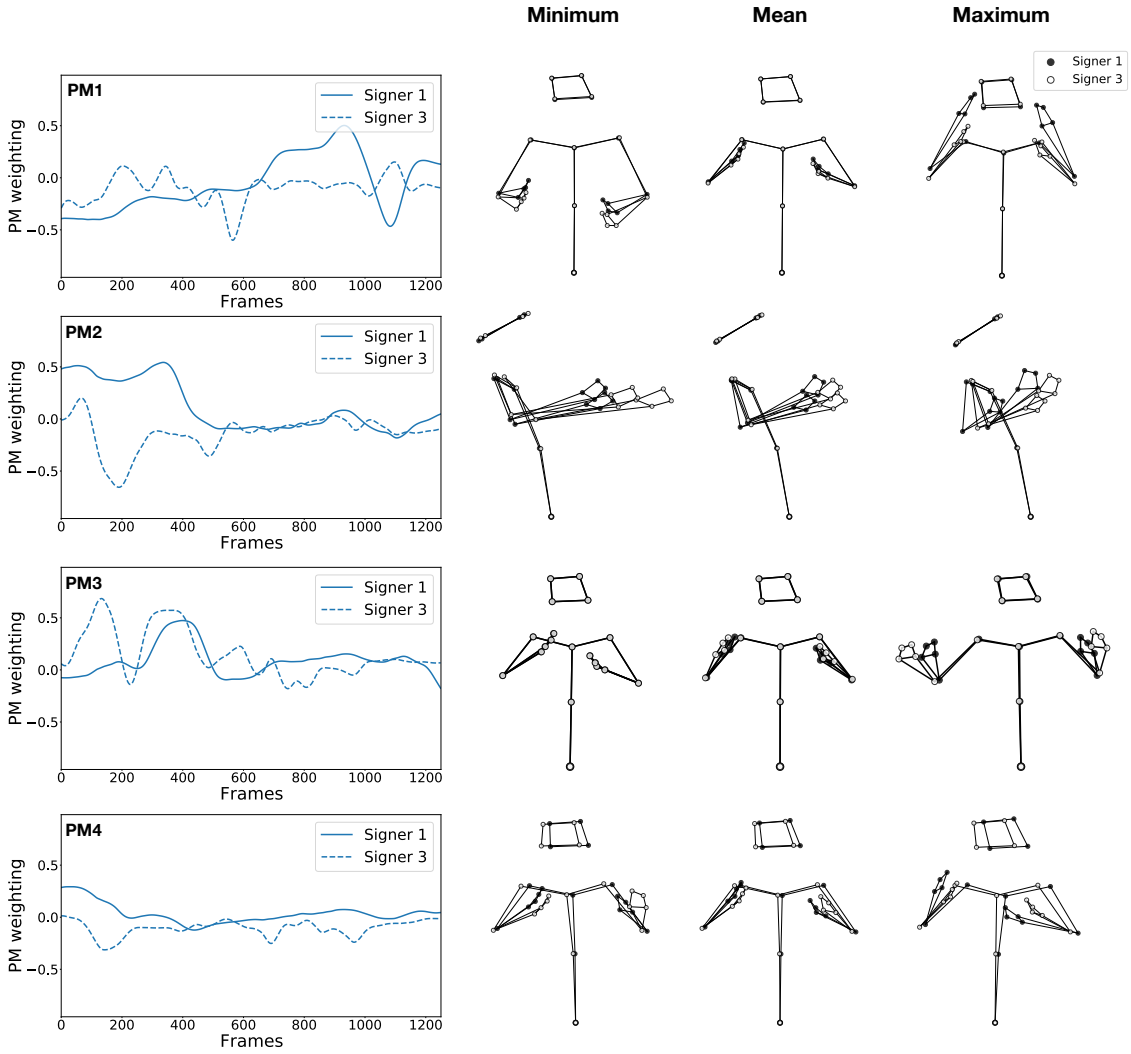


FIGURE 7.3: The first four common PMs, for Signers 1 and 3. Left: weightings. Right: stick figures of important postures during the PM (minimum, maximum and mean of the signer's first mocap example, along the direction of the PM).

7.3.3 Individual Principal Movements compared with common Principal Movements

As shown in Figure 7.4, the first eight individual PMs explained most of the variance in the mocap data of individual signers. Combined, they explained 95.9% (SD=0.6%) of the cumulative variance. By comparison, the first eight common PMs explained 94.9%, and the first seven individual PMs explained 94.6% (SD=0.8%). Therefore, although the common dataset contained unsynchronized movements performed by six different signers, the common PMs explained a similar amount of the movements variance, compared with PMs computed separately for each signer.

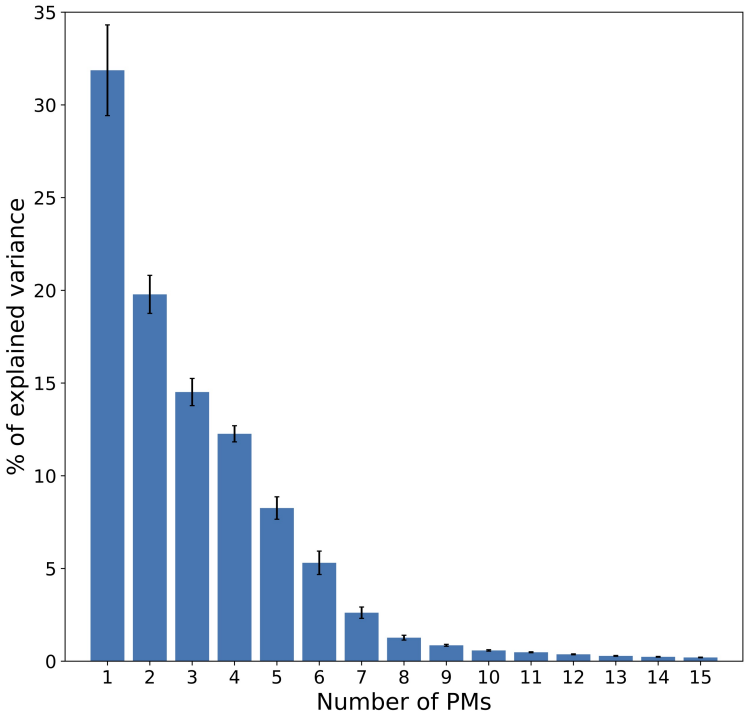


FIGURE 7.4: Variance explained by the first 15 individual PMs. Error bars indicate standard errors across signers.

The first eight individual PMs of Signer 2 (Videos 7.15 to 7.22) are shown in Figure 7.5. Most of the individual PMs were similar to the common PMs, although they were sometimes ranked in a different order. For instance, PM1 was characterized as vertical parallel movements of the hands, jointly with anteroposterior parallel movements of the hands. This first individual PM was thus a combination of common PM1 and PM2. PM2 was characterized as mediolateral opposite movements of the hands (common PM3), PM4 as vertical parallel movements of the hands (common PM1), PM5 as mediolateral parallel shift of the hands (common PM5), PM7 as shoulder abduction jointly with head tilt (common PM8) and PM8 as wrists flexion jointly with upper-body rotation around the vertical axis (common PM7). Other individual PMs were a combination of a similar common PM with additional motion patterns. For instance, PM6 quantified vertical opposite movements of the two hands (common PM6), but combined with lateral sway of the upper-body. PM3 quantified upper-body rotation around the vertical axis (common PM4), but combined with mediolateral movement of the left hand toward the right hand. Individual PM3 could also be considered as specific to Signer 2, because this mediolateral movement of the left hand only did not occur in the common PMs, and the upper-body rotation

it quantified was not as wide as in common PM4. Moreover, the mediolateral movements of the hands in PM2 and PM3 of Signer 2 were mainly related to one specific hand. For instance, the right hand had a higher movement amplitude in PM2, while the left hand had a higher amplitude in PM3. Thus, common PM3 can be seen as the combination of PM2 and PM3 of Signer 2.

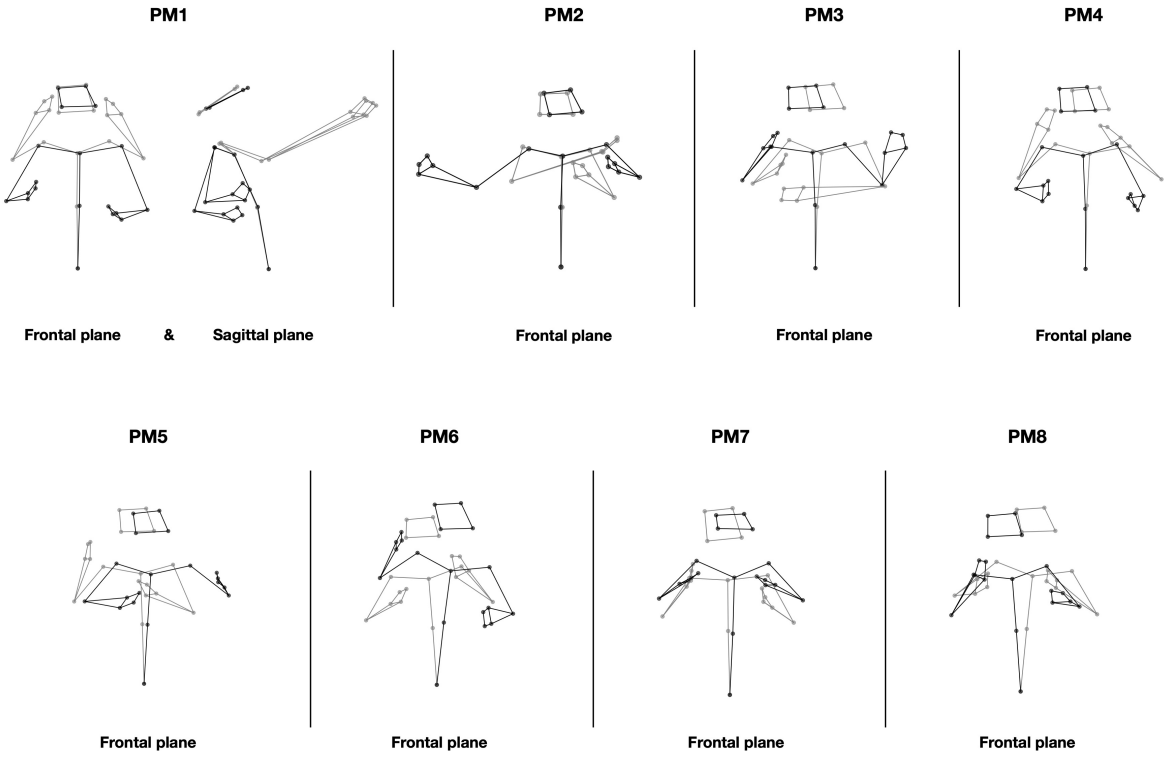


FIGURE 7.5: The first eight individual PMs of Signer 2. Stick figures represent the PM at the time instants corresponding to the minimum (gray) and maximum (black) PM weighting, across the 24 examples of the signer. PMs are displayed in their main plane of motion (e.g., frontal and/or sagittal).

These observations can be extended to the other signers. For instance, the characteristic patterns of common PM1 similarly occurred in PM1 of Signer 3 (Video 7.23) or in PM1 of Signer 5 (Video 7.24). PM1 of Signer 1 (Video 7.25) was a combination of common PM1 and PM3. Similarly, PM3 of Signer 1 was a combination of common PM2 and PM3 (Video 7.26). In summary, most of the individual PMs were similar to the common ones. They were sometimes performed differently from one signer to another, or were combinations of several common PMs. A few signer-specific PMs were found. Still, most of them could be interpreted as signer's specific execution of common PMs.

7.4 Discussion

In the present chapter, we used PCA to decompose the mocap data of spontaneous LSF produced by six signers into a reduced set of simpler, elementary, movements. The original motion data were transformed into a new space spanned by principal components, called principal movements (PMs). The first eight individual PMs (i.e., computed separately from each signer's mocap data) explained 95.9% (SD=0.6%) of

the variance in the movements of individual signers. To the author's knowledge, this is one of the highest numbers of PMs needed to account for such amounts of variance, compared with other similar motion data. Troje (2002a) has reported that 98% of the variance in walking patterns was covered by the first four PMs. The first four PMs also covered 98% of the overall variance in diving (Young and Reinkensmeyer, 2014) and 95.5% of the variance in skiing (Federolf et al., 2014). For juggling, even in the hardest conditions (5-balls handling), six PMs were sufficient to explain over 96% of the variance (Zago et al., 2017b). This result supports SL as a complex set of movements involving multiple body parts and various semantic structures. A higher number of PMs have been reported for the finger gestures of pianists, whose first four PMs only explained 60% of variance in Furuya et al. (2011) and which required eight to ten PMs to explain 95% of the overall variance in Tits et al. (2015). Similarly, six to nine PMs were needed to explain 95% of the variance of finger gestures during hand manipulation, such as grasping, holding keys or reading a book (Todorov and Ghahramani, 2004). As regards SL motion, 95% of variance in the finger gestures of ASL alphabetical signs was accounted by 11 PMs in Yan et al. (2020). Therefore, finger gestures may involve higher dimensions than movements of other body parts in motor control, including for SL production. However, in both finger gestures (Yan et al., 2020) and upper-body movements (the present study), SL seems to require a higher number of motor components, compared to non-SL movements.

Still, despite the complexity of SL, the number of PMs reported in the present study remains relatively low. Moreover, a similarly low number of common PMs (i.e., computed on the whole dataset containing all signers) accounted for 94.9% of the overall variance. This is a particularly intriguing result as the analyzed movements were not synchronized in time across signers and examples. Indeed, in the present study, each signer freely described pictures in LSF, without being required to use specific gestures. By contrast, most of the studies mentioned above have investigated movement sequences that shared similar temporal structures (e.g., gaits (Troje, 2002a), karate's *kata* (Zago et al., 2017a), dives (Young and Reinkensmeyer, 2014), juggling patterns (Zago et al., 2017b) or isolated ASL alphabetical and number signs (Yan et al., 2020)). In Tits et al. (2015), finger gestures were potentially not synchronized across pianists and excerpts. However, PMs were extracted separately on each excerpt, which may have reduced the variability of the analyzed movements.

Moreover, the amount of variance explained was very similar between common and individual PMs, suggesting a common structure of SL movements shared across signers. The distinction between individual and common PMs was inspired by the prior work of Federolf et al. (2014) on skiing. By comparison with our study, the PMs common to all skiers did not represent the original skiing movements as well as the PMs computed for each skier. The first four common PMs explained 88% of the variance, compared with 95.5% (SD=0.5%) for individual PMs. Again, signers in our study freely produced movements to describe the different pictures, while skiiers in Federolf et al. (2014) executed trials under the same race conditions. This outcome suggests that SL motion presents less variablity across individuals and examples, compared with other movements.

The motion patterns described by common PMs were mostly visible in the frontal and sagittal planes, in line with the vast majority of prior studies (Troje, 2002a; Federolf et al., 2013b; Federolf et al., 2013a; Federolf et al., 2014; Federolf, 2016; Zago et al., 2017a; Zago et al., 2017c; Promsri and Federolf, 2020). However, Zago et al. (2017b) have reported important PMs of juggling, such as trunk rotation or upper-limbs internal flexion and extension, in the transverse plane. In the present study, common PM4 (i.e., trunk rotation) can be described in the transverse plane

but it was clearly visible from a frontal view. PM2 was best visible from a sagittal view. Still, the patterns quantified by most PMs were occurring in the frontal plane. Similarly to prior work, first PMs quantified simple movements, while higher-order PMs extracted finer motion descriptors, such as shoulder abduction or wrist flexion. Some motion patterns seemed to be specific to SL, compared with previously studied movements. For instance, a high number of PMs quantified hand movements and some PMs outlined specific covarying movements between the head and the arms. Further work examinating mocap data of fingers or facial expressions could be of interest, in order to extend these first observations. Testing PM decomposition on larger datasets could also allow for some generalization, beyond specifities due to the linguistic context or to signer-specific occasional movements.

In addition to the similar amounts of variance they explained, individual and common PMs described highly similar motion patterns. Some individual PMs could be considered as signer-specific but most of them quantified the same movements as common PMs, either separate or combined. These inter-individual observations are in line with the ones made in skiing (Federolf et al., 2014). Unlike Tits et al. (2015) or Zago et al. (2017b), no major inter-individual differences were found in the number of PMs needed to account for high variance in the movements. This may reflect that the six signers of this study presented a similar level of expertise in SL gestures, all being fluent signers. However, common PMs were sometimes executed differently by signers, as shown for skiing (Federolf et al., 2014). For instance, PM2 revealed more remote position of the hands from the body for Signer 3, compared with Signer 1. During the hands movements of PM3, signers also differently spread their hands apart. Moreover, PM4 outlined that the upper-body rotation of Signer 1 was mainly leftward, while the one of Signer 3 was mainly rightward.

Taken together, these results suggest that SL motion has a common structure across signers and discourses that can be decomposed into a reduced set of elementary movements. This specificity, compared to priorly studied movements, such as walking (Federolf et al., 2013a; Troje, 2002a), may be explained by the fact that not only are SLs constrained by biomechanical rules but they also follow a highly-structured linguistic system, which is shared among signers. Taking inspiration from prior research in sport science, gait analysis and postural control, the application of PCA has several advantages and opens up promising perspectives for research in SL. Overall, this decomposition method is of great interest as it allows extracting motion patterns from a holistic perspective, which can be considered as analogous to how human observers may quantify movements (Federolf et al., 2014). More specifically, its potential contributions to SL are two-fold. First, its invertibility allows resynthesizing PMs in the original 3D space. Researchers can gain insights into the complex structure of SL movements, by visualizing these directions of high movement variability and by assessing inter-individual differences. For these reasons, PM decomposition has been widely used to evaluate gesture expertise and could be of interest for SL problems, such as motor learning (Zago et al., 2017b; Zago et al., 2017a; Young and Reinkensmeyer, 2014). Second, it allows for dimensionality reduction. A few studies previously aimed at drastically reducing the frame rate (Foulds, 2004) or the number of markers (Tartter and Knowlton, 1981) in SL telecommunication, for bandwidth reasons. Following the present results, dense mocap datasets could be considerably reduced using only a subset of PMs while keeping most of the information. This outcome calls for further work evaluating the observers' comprehension of SL messages when resynthesized from the PMs. Furthermore, both the potential to resynthesize movements from the PMs and the potential to reduce dimensionality make PM decomposition very promising for the improvement of technologies

for SL automatic processing. For instance, it could ease the incorporation of high-dimensional mocap recordings in SL movement generation and, thus, break down barriers caused by the lack of naturality and comprehensibility of virtual signers.

Part II Summary

In real-life conditions, Sign Language (SL) is a complex and continuous stream of motion features from various body parts, which makes it challenging to model. The main objective of this part of the thesis was to provide novel representations and processing tools for modelling and analyzing spontaneous SL motion. For that aim, we used the MOCAP1 corpus, which provides 3D motion capture (mocap) data of spontaneous French Sign Language (LSF) utterances produced by six different signers. Using this corpus, preprocessing methods were developed, including the normalization of mocap data with respect to the size, shape and posture of the signers. We then showed that the frequency content of spontaneous LSF movements could be properly described using a 0–12-Hz bandwidth. This range is wider than prior estimations made with isolated signs but is still drastically lower than the raw spectral information provided by most mocap systems. Furthermore, although spontaneous SL involves a wide variety of motion patterns and linguistic forms, a Principal Component Analysis (PCA) applied to our mocap data revealed that the LSF discourses could be described by a reduced set of eight simple, one-directional, Principal Movements (PMs), across signers and linguistic contents. In summary, despite the intrinsic complexity of spontaneous SL motion, the actual information needed to analyze it seems to lie within spaces of lower dimensions (e.g., in low frequencies or low-order PMs). Could these motion representations (or further ones) be used to automatically identify signers during SL discourses, as previously shown for non-SL movements (see Chapter 4)? We aim to answer this question in Part III.

Part III

Person identification from motion: the case of Sign Language

Chapter 8

Identity information in the movements: insights from human perception

Humans observers can identify individuals from biological movements, such as walking or dancing (see Section 2.1.3). It remains to be investigated whether SL motion, which obeys to linguistic constraints in addition to biomechanical ones, also allows for person identification. The study of the present chapter is the first to investigate whether deaf perceivers actually identify signers based on mocap data only, using Point-Light Displays (Section 8.1). Further computational analyses of the mocap data were then conducted in order to test the role of morphological differences between signers in the identification (Section 8.2). The present behavioral and computational findings suggest that mocap data contain sufficient information to identify signers, beyond simple cues related to morphology. These results form the basis for the further machine learning developments of this thesis (see Chapter 9). Indeed, they scientifically demonstrate that human observers can identify signers from their movements, which confirms the need for novel technological tools able to control identity-specific aspects of SL motion, in particular for anonymization purposes. Moreover, the minor role of morphology-related cues in the human ability to identify the signers calls for further work, including machine learning studies, investigating the role of other motion features, in particular kinematic ones (Section 8.3).

This chapter is partly reproduced from Bigand et al. (2020).

8.1 Human ability to identify signers from mocap data

The present study evaluated the ability of deaf perceivers to identify signing individuals that were presented as Point-Light Displays (PLDs). PLDs were computed from a subset of MOCAP1-v2 corpus (see Chapter 5) in which four different signers freely described pictures in French Sign Language (LSF). The specific four signers were chosen because of their different levels of exposure to the general public in deaf communities. With these different levels of exposure, we hypothesized that the participants may report different degrees of familiarity with the signers. The familiarity with each signer was reported by the participants at the beginning of the experiment. Then, short excerpts of the LSF descriptions were randomly presented to the participants. For each excerpt, participants were asked to identify the signer with a four-alternative forced choice (Section 8.1.1). Depending on familiarity level with participants, some signers were identified with substantial accuracy (Section

8.1.2), which is discussed as compared to prior work on non-SL movements (Section 8.1.3).

8.1.1 Methods

Participants

24 participants (mean age = 42.0, SD of age = 11.0) took part in the study. The Research Ethics Committee of Paris-Saclay University validated the experiment. All participants were deaf and were users of French Sign Language. Language level was self-assessed in the beginning of the experiment. Participants mainly reported the highest levels C1-C2[1] (79,17%). 12% reported advanced levels (B1-B2). 8.33% reported intermediate level A2.

Stimuli

PLDs were generated, similarly to the original idea of Johansson (1973). Major joints of the body were displayed with white dots on a black background, based on the 3D motion data of MOCAP1-v2 (see Chapter 5). From this dataset, we selected 16 different descriptions performed by the first four signers and trimmed them regardless of the linguistic content. The four signers were chosen so that we could expect familiarity with signers to be gradual (see Section 8.1.2). All markers were presented as global positions in a reference system with the pelvis as the origin (see Figure 8.1). All the stimuli were displayed in front view (see Video 8.1).

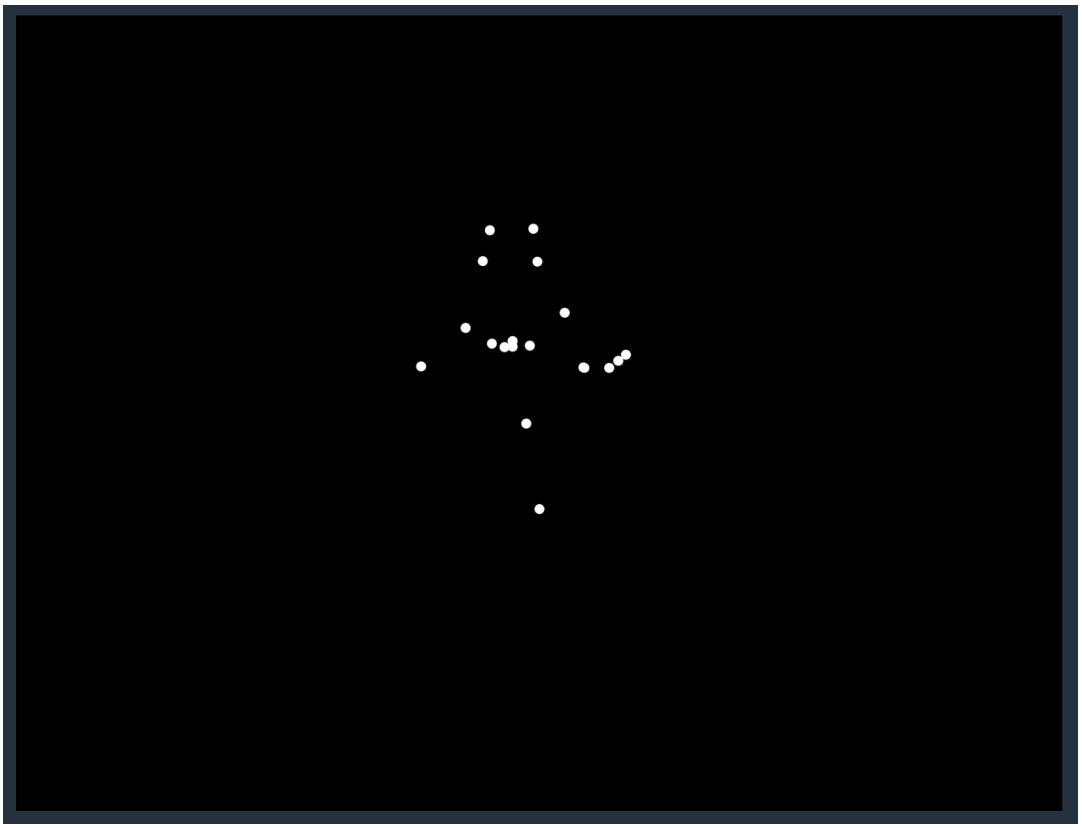


FIGURE 8.1: Example of the Point-Light Displays (PLDs) used in the experiment (all in front view).

[1] European CECRL levels

Design and procedure

The participants took part in the experiment via an online survey. The mean duration of the survey across participants was 11.04 min. Most participants used a computer (70.83%) rather than a tablet or smartphone.

First, the familiarity of the participants with the four signers was expected to be variable because of their different exposure to deaf people. Signer 1 has been a popular story-teller for children and producer for deaf TV shows, being the first deaf person seen on TV in France, in 1979. Signer 2 is an LSF translator and actor. He notably worked as a translator for Websourd, a highly popular deaf media. Signer 3 is a young LSF journalist working for different media. We were expecting that she would get lower recognition rates as she only recently appeared in the field. Signer 4 is involved in some projects on LSF but with lower exposure. She has worked as an LSF translator and trainer but mainly in a local environment. To verify these background differences, before the test session, all of the four signers were presented on the screen with their names and three photos taken from public content (TV, Youtube, conferences...) (see Figure 8.2). Participants were then asked to report their familiarity with each signer by answering this question : *"Have you ever seen this person?"*. Four levels could be reported: *"No, never"* (0), *"Yes, occasionally"* (1), *"Yes, often"* (2) and *"Yes, very regularly"* (3).

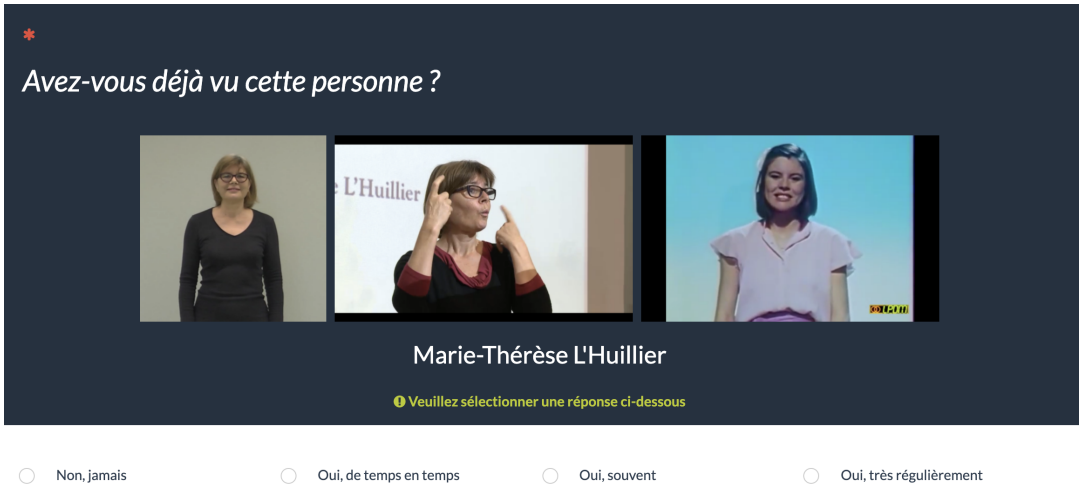


FIGURE 8.2: Example of the presentation of the signers (here Signer 1) and familiarity evaluation. Participants were asked to report their familiarity with each signer on a four-level scale: *"Have you ever seen this person?"*. Possible answers were as follows: *"No, never"* (0), *"Yes, occasionally"* (1), *"Yes, often"* (2) and *"Yes, very regularly"* (3).

After that, the test session consisted of 16 trials (4 signers x 4 picture descriptions). In each trial, the Point-Light Display (PLD) (mean duration = 10.8s, SD = 2.6) was presented, followed by the presentation of four buttons, illustrated by each signer's photo (see Figure 8.3). The PLD videos were launched automatically, played only once and participants could neither pause the video nor rewind it. When the video was finished, the survey automatically displayed the four choices shown in Figure 8.3. Then, participants were asked to identify the signer in this four-alternative forced choice where each choice was one of the four signers. All of the 16 stimuli were presented in random order. No response feedback was given to the participants. All instructions were given in both written French and French Sign Language using pre-recorded videos (see Figure 8.4).

FIGURE 8.3: Example of the four-alternative forced choice presented after the moving Point-Light Display (PLD) example: *"Please select the person you have recognized"*. Each choice was one of the four signers, each illustrated by their photo.

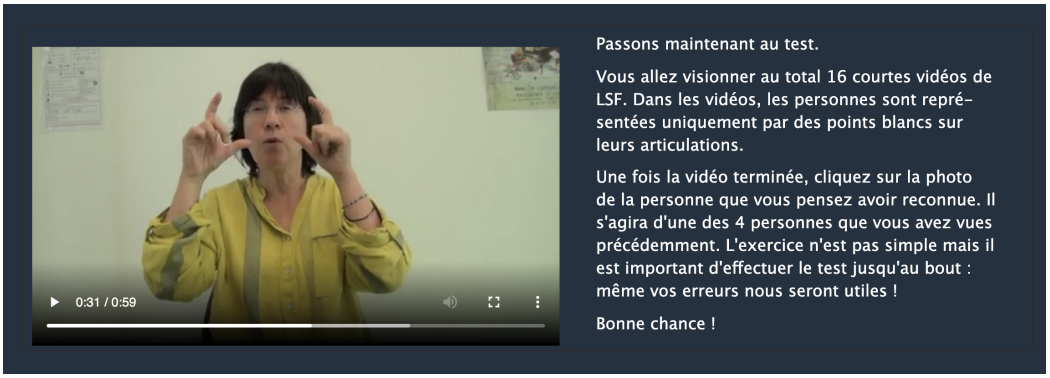


FIGURE 8.4: Example of instructions provided in written French and French Sign Language (here before the test session): *"Let's take the test. You will watch a total of 16 short videos of LSF. In the videos, the people are represented only by white dots on their body joints. When the video is over, please click on the picture of the person you think you recognized. It will be one of the four people you saw earlier. The task is not easy, but it is important to complete the test: even your mistakes will help us. Good luck!"*.

8.1.2 Results

A repeated-measure one-way ANOVA was performed with signer (four levels) as within-subjects factor and self-reported familiarity as dependent variable. A significant main effect of signer was found on self-reported familiarity ($F(3,69) = 6.65, p < .001, \eta^2 = .13$). Bonferroni-adjusted post-hoc tests revealed that familiarity was significantly lower for Signer 4 (M=0.96) than for Signer 1 (M=1.75, $p < .01$), Signer 2 (M=1.67, $p < .01$) and Signer 3 (M=1.71, $p < .05$). No significant differences were found between the 3 first signers. Self-reported familiarity levels as a function of the four signers are shown in Figure 8.5.

Then, a repeated-measure one-way ANOVA was performed with signer (four levels) as within-subjects factor and correct identification as dependent variable. A significant main effect of signer was found on correct identification ($F(3,69) = 12.36, p < .001, \eta^2 = .25$). Bonferroni-adjusted post-hoc tests were performed to test for differences between signers. They revealed a significant increase ($p < .001$) in performance between Signer 4 (M=30.2%) and Signer 1 (M=65.6%). There was an increase between all four signers (30.2%, 36.5%, 45.8%, 65.6%) but no significant differences were found between Signer 4 and both Signers 2 and 3.

One sample Student's t-tests revealed that identification performance was significantly above chance level (25%) for Signer 1 ($t(23) = 8.21, p < .001$), Signer 2 ($t(23) = 4.05, p < .001$), Signer 3 ($t(23) = 2.30, p < .05$) but not for Signer 4 ($t(23) = 1.16, p = .26$). Identification scores as a function of the four signers are shown in Figure 8.6.

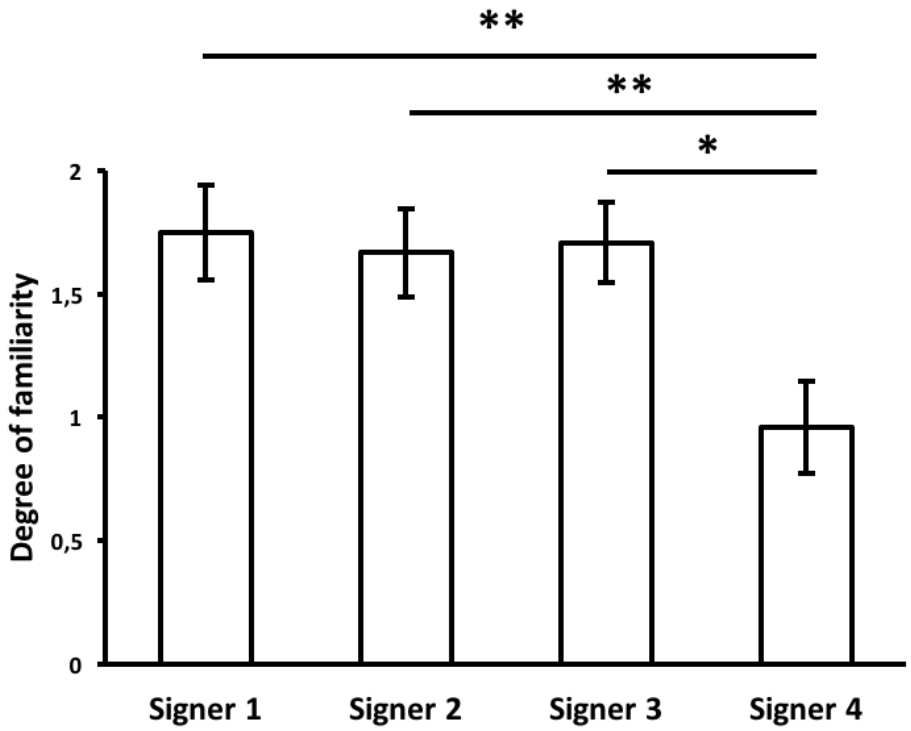


FIGURE 8.5: Self-reported familiarity for each signer, averaged over participants. Error bars indicate standard errors. Significant differences between signers : *($p < .05$), ** ($p < .01$).

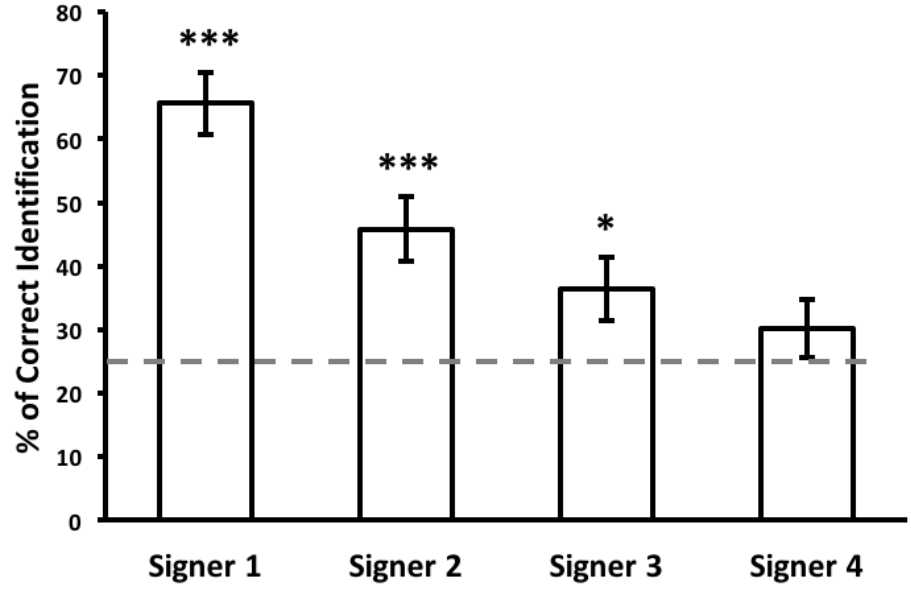


FIGURE 8.6: Performance scores from the four-alternative forced choice identification task, averaged over participants. Dashed horizontal line indicates the chance performance level (i.e., 25%). Error bars indicate standard errors. Significant differences from chance level : * ($p < .05$), *** ($p < .001$).

8.1.3 Discussion

Although it is not an easy task for participants to evaluate the familiarity with signers, the analysis revealed significant differences, distinguishing Signer 4 from the three other signers. This finding provides a partial account for the correct identification scores, which differed from chance level for all signers except for Signer 4. This lower value could be explained by the lower exposure of Signer 4 to the general public. In other words, participants managed to identify familiar signers above chance level. It suggests that the mocap data we used in the experiment included sufficient information for participants to identify familiar signers. Further work would be needed to explain the differences in correct identification between the three first signers, despite equal self-reported familiarity.

The first study by Kozlowski and Cutting (1977) has reported performance scores above chance level (16%) but only reaching 38%. Troje et al. (2005) have reported 76% correct identifications but involving extensive pretraining for the participants. Our results (Figure 8.6) included the responses of all participants, whatever their familiarity with signers. In addition, limitations of the online survey could be discussed. The design of the survey ensured that participants were not able to pause the video or to stop it before the end, but the conditions under which each participant responded to the survey could still vary. Therefore, reaching identification scores such as 65.6% or 45.8% for some signers suggest that their movements provided critical information for identification. Further computational analyses of the mocap data then allowed us to better understand the nature of this information, in particular as regards morphology of the signers.

8.2 The role of morphology in the identification

The excerpts that were presented to the participants were rather short (mean duration = 10.8s, SD = 2.6) and were randomly extracted from the original recordings, regardless of the linguistic content. The data we used neither did include facial nor finger markers. Prior studies have demonstrated that a precise display of the fingertips was needed to ensure comprehensibility in SL, especially for lexical signs (Poizner et al., 1981). Other studies have pointed out the crucial role of facial components during the comprehension of ASL (Emmorey et al., 2009), in particular mouthing and eye gaze. None of these informations was present in the PLDs presented here to the participants. It is therefore assumed that the identification of signers was achieved beyond simple differences in linguistics. One aspect of motion information present in the PLDs that we thought important to test is morphology. Based on data-driven methods, we defined a morphology factor that optimally described morphological differences between the four signers (Section 8.2.1). We then assessed the extent to which this factor was correlated with the participants' responses in the visual perception experiment presented in Section 8.1 (Section 8.2.2). These analyses revealed that, beyond morphology, further motion features needs to be investigated to account for the human ability to identify the signers (Section 8.2.3).

8.2.1 Morphology: a PCA-based definition

As presented in section 8.1.1, skeletons of the four signers were displayed as global coordinates for which the pelvis was the origin.

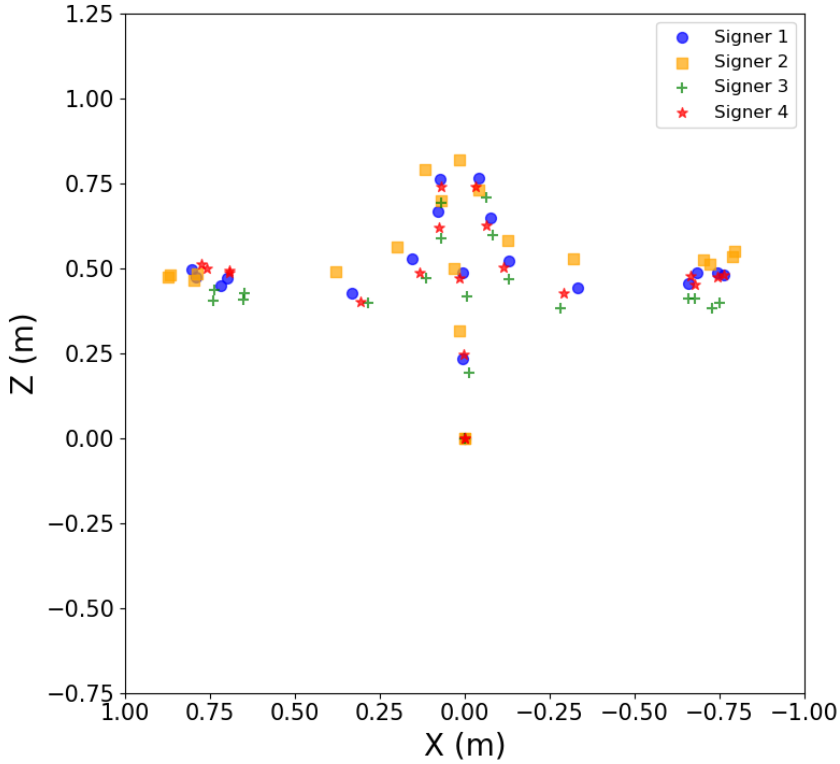


FIGURE 8.7: Reference posture of the four signers.

This coding allowed for consistent comparison of the signers who had been placed and oriented differently in the motion capture space. Nevertheless, different morphologies of the four signers were represented in this coordinate system, as shown in Figure 8.7. To evaluate the role of these differences in the identification, we used a PCA-based approach in order to compute morphological similarities and to compare it with participants' confusion errors between signers.

Our aim was to assess the extent to which morphology could account for participants' performance in the experiment. As it can be defined with various indices such as height or shoulder width, we ran a PCA to find the most relevant variables to represent morphology. Similarly to Tits (2018), PCA was performed on distance from head to pelvis, distance from hand to hand (in extension), shoulder width, and individual segment lengths (trunk, arm and forearm).

The first principal component accounted for 72% of variance in the data, and it was highly correlated with the distance from head to pelvis ($r(2) = .99, p < .05$). Consequently, this variable was chosen to define morphology. Figure 8.8 illustrates morphological differences between all signers, using this factor. This emphasizes an important gap between Signer 2 (highest) and Signer 3 (lowest) values, and specifies a higher value for Signer 1 than for Signer 4, although both fit in a similar range.

8.2.2 Influence on participants' responses

Based on the morphology factor defined in the previous section, a similarity matrix was computed among all signers. Using the Euclidean distance, similarity was computed as follows :

$$s_{i,j} = \frac{1}{1 + \sqrt{(m_i - m_j)^2}} \tag{8.1}$$

where m_i is the normalized morphology factor of Signer i.

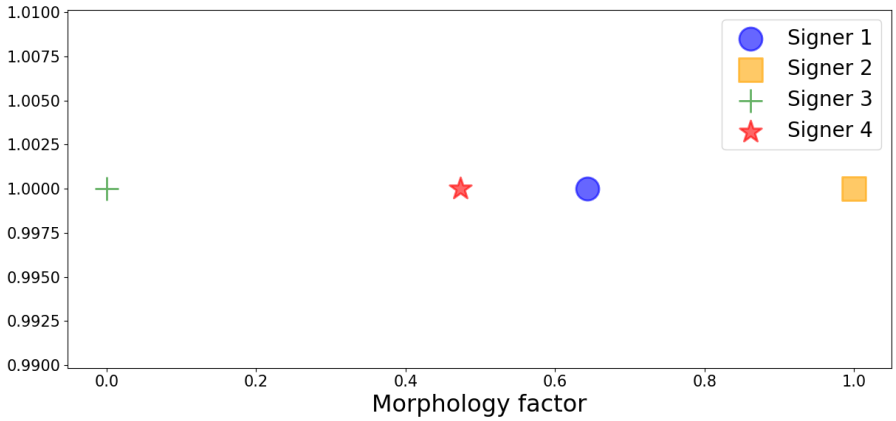


FIGURE 8.8: Ranking of the four signers as a function of the normalized morphology factor.

The computed similarity matrix is shown in Figure 8.9 (left). Each row represents the amount of similarity of a signer with the three other signers. According to this representation, Signer 2 is more likely to be confused with Signer 1 (39%) than with Signer 3 (26%). Signer 4 would have equal chances to be confused with both Signers 2 and 3. This measure based on computational and statistical analyses of the mocap data allowed us to predict confusions related to morphological cues in the identification of signers. We compared the similarity measures to the actual confusions of participants in the experiment, across the four signers. As shown in Figure 8.9, 43% of the confusions for Signer 2 are made with Signer 3, while the latter signer reported the lowest similarity measure (26%) with Signer 2. In the case of these two signers, the less morphologically similar they are, the more confused they seem to be. This suggests that morphology may not be the main information used by participants to identify the signers.

The Pearson's correlation between the similarity matrix and the confusion matrix was measured. The resulting coefficient revealed that correlation is not significant ($r(10) = -.36, p = .26$). Taking into account only familiar signers (i.e., the three first signers, see Figure 8.5), correlation between morphological similarities and participants' confusions is also not significant ($r(7) = -.20, p = .61$).

FIGURE 8.9: Morphological similarity among signers (left). Participants confusions between signers (right).

8.2.3 Discussion

As previously detailed in Section 3.2.1, the role of morphology in motion perception has been a matter of debate for several types of movements. Sie et al. (2014) have proposed a simple skeleton scaling method, by placing the coordinate system on a reference node of the body (i.e., on the pelvis), and dividing all node coordinates by the torso height. Troje et al. (2005) have used normalizations with respect to size or/and shape, using linear regression models. In the specific context of LSF motion, our computational analysis based on PCA revealed that morphological similarities between signers were not correlated with participants' confusions. This suggests that morphology alone cannot account for correct signer identifications and calls for further investigations of other candidate features for identification.

8.3 Further insights from machine learning: preliminary observations

The visual perception experiment of the present study (Section 8.1) provides the first evidence that deaf perceivers managed to identify familiar signers, shown as PLDs, above chance level, as demonstrated for walking (Troje et al., 2005) or dancing (Loula et al., 2005). The second outcome of the study is that morphology was not sufficient to identify the signers (Section 8.2). A computational analysis based on PCA revealed a non-significant correlation between the participants' confusion errors and the morphological similarities among signers. This is also consistent with prior studies on the perception of identity from gait (Troje et al., 2005). In the latter study, even after having removed size and shape information, the different walkers were still identified with high accuracy (i.e., about five to six times higher than chance level).

Combining human data and computational analyses, the main findings of the present study suggest that SL mocap data contain enough information for signers to be identified, and this beyond morphology-related cues. Given that the present PLDs were as short as 10 seconds, randomly selected in the original recordings and as it was demonstrated that fingers were needed for SL comprehensibility (Poizner et al., 1981), linguistics were unlikely to play a major role in the identification. Participants thus may have used other cues, such as kinematic cues in particular. To address this question, we aimed to develop a machine learning model for automatic signer identification.

Before the extensive developments of the machine learning model (Chapter 9), preliminary analyses of the mocap data of the four signers used in the human experiment were conducted. The aim of these preliminary tests was to assess the extent to which kinematic features could be used by machine learning models to distinguish between signers from mocap data. From each of the 24 mocap recordings, a 10-sec mocap excerpt was extracted. Each mocap recording was then represented as a $(1 \times 142,500)$ flattened vector (i.e., 3D coordinates of 19 markers during 10 secs, sampled at 250 fps). PCA was applied to the matrix containing the 24 mocap vectors of all signers. Using the computational methods presented in Section 5.3.2, the mocap vectors could be normalized with respect to size and shape (shape-normalized data being also size-normalized). Preliminary observations were made on the first two PCs extracted by the PCA. Without any normalization (ORI), the first two PCs explained 38% of the variance in the original vectors and also accounted for some

differentiation between signers, as shown in Figure 8.10. We also noted that the different mocap examples of each signer were consistently distributed across the two axes. The first PC allowed differentiating Signer 2 from all other signers, while the second PC allowed differentiating Signer 3 from Signers 1 and 4. Interestingly, these PC projections highly reflect the ranking of the four signers as a function of morphology (Figure 8.8), which suggests that the first PCs captured critical information about the morphology-related cues of the signers.

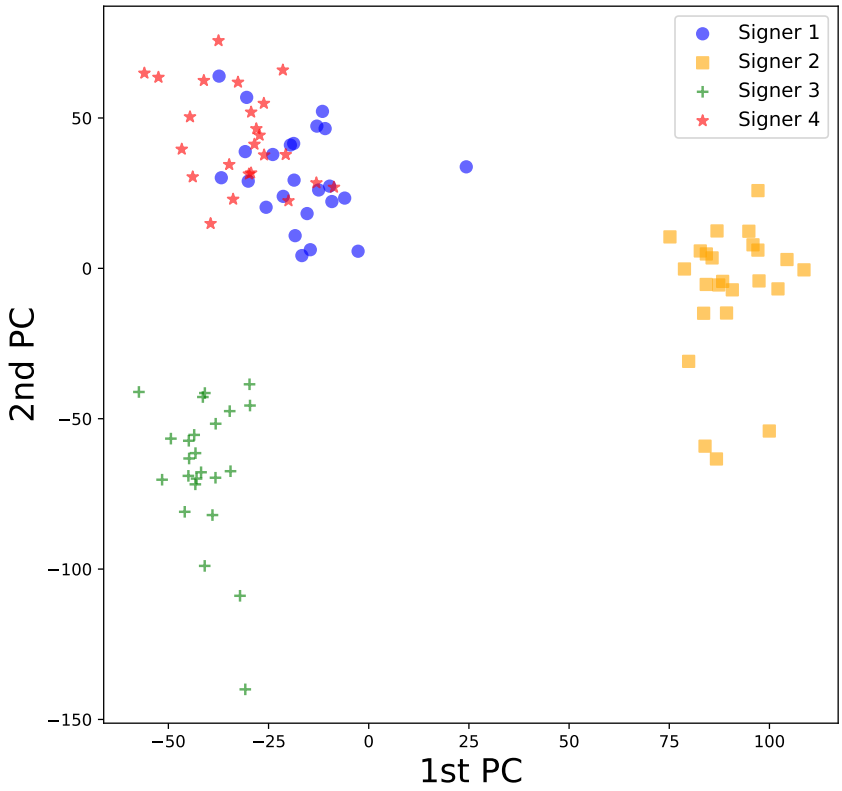


FIGURE 8.10: Projections of the original (ORI) mocap data onto the first two PCs.

When the mocap data were size- and shape-normalized, the first two PCs accounted for a lower amount of variance (27%) in the mocap vectors than for original motion but they still allowed for some differentiation between signers, as shown in Figure 8.11. The first component discriminated between Signer 3 and Signer 4, while the second component discriminated between Signer 1 and all other signers. The lower percentage of information given by these two components was expected as data were normalized. The main outcome of these preliminary results is that PCA still extracted substantial information that discriminates between signers, even when size and shape information were removed.

Although the mocap data analyses conducted in this preliminary step were quite simple, the observations made on PC1 and PC2 support the possibility to automatically identify signers from mocap data, beyond cues related to anthropometric differences (e.g., in body size or shape). The role of cues related to the posture of the signers was not the focus of this study. It will be further assessed in Chapter 9. Beyond postural information, as motivated by prior evidence from both human visual perception (Troje et al., 2005; Westhoff and Troje, 2007) and machine learning models (Zhang and Troje, 2005; Troje, 2002a; Carlson et al., 2020), kinematics may

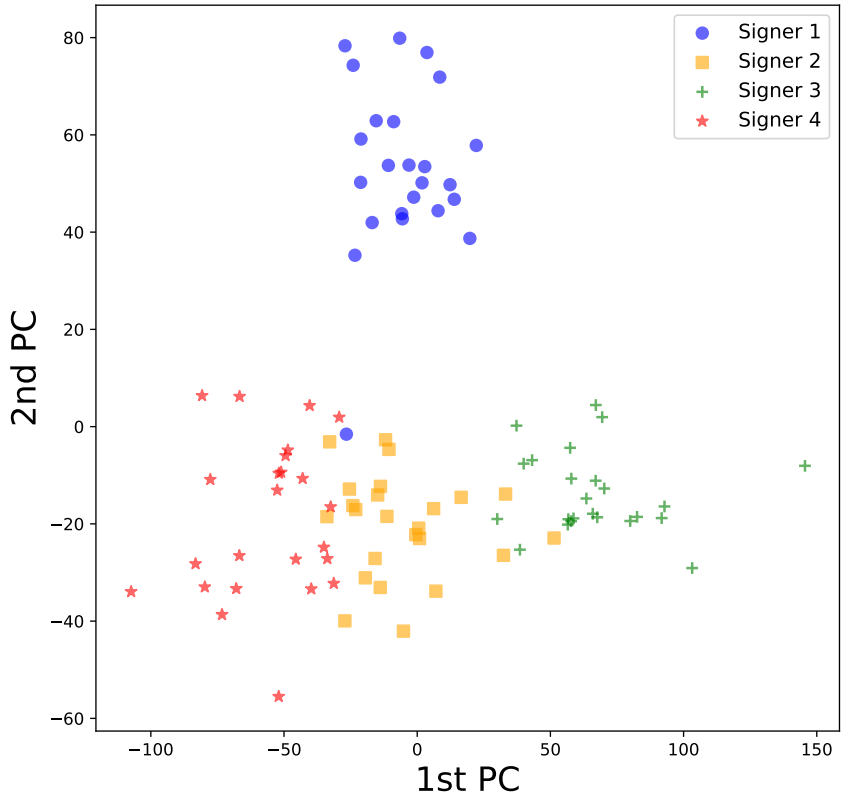


FIGURE 8.11: Projection of the shape-normalized (SH) mocap data onto the first two PCs.

play a major role in how observers extract human attributes from motion. Taken together, these outcomes call for additional research further investigating the motion cues that allow for signer identification, in particular kinematic ones. Based on the prior work presented in Chapter 4, an original machine learning model aimed to automatically determine the identity-specific features of signers is proposed in the following chapter.

Chapter 9

Machine learning of motion reveals the kinematic signature of identity

As shown in Chapter 8, SL motion contains information about the identity of a signer, as does voice for a speaker or gait for a walker (see Section 2.1.3). However, how such information is encoded in the movements of the signers remains unclear. In the present chapter, a machine learning model was trained to extract the motion features allowing for the automatic identification of six signers, based on the mocap data of MOCAP1-v2 (see Chapter 5 for corpus description) (Section 9.1). The performance of the model on original, size-, shape- and posture-normalized mocap data further confirmed that the identity of a signer can be conveyed by kinematics alone. The further discriminant statistics used by the model in the identification defined the kinematic signature of the identity of signers (Section 9.2). These findings constitute a first step toward determining the motion descriptors necessary to account for the human ability to identify signers (Section 9.3).

This chapter is partly reproduced from Bigand et al. (2021c).

9.1 Methods

Based on various motivations, notably related to the time-invariant property of identity and the structure of spontaneous LSF motion, a statistical-based approach was taken to model the mocap data of the signers of MOCAP1-v2 (Section 9.1.1). PCA, followed by a linear classifier was used to automatically identify the six signers (Section 9.1.2). This machine learning model was then trained and tested using cross-validation across the 24 mocap examples. Finally, the discriminant statistics used by the classifier to identify the signers were analyzed (Section 9.1.3).

9.1.1 Motion model: a statistical-based approach

Similarly to Chapter 7, 24 mocap examples of 5-second duration each were extracted from the original recordings of MOCAP1-v2 (see Chapter 5) and low-pass filtered using a 4th-order Butterworth Filter with a cutoff frequency of 12 Hz. The start frame of each mocap example was fixed after the initial "T" posture of the signers. The end frame was then deduced from the 5-second duration, irrespective of the semantic content in the original recording. For each signer, each mocap example was related to a different picture description in LSF. This resulted in 24 mocap examples per signer, of 5-second duration each (see examples in Videos 7.1 to 7.6). Using the computational methods presented in Section 5.3.2, all mocap examples could be kept as original (ORI) or normalized with respect to size (SI), shape (SH) and

posture (POST) of the signers. As a reminder, size, shape and posture of the signers are structural cues (see Section 5.3.2), by contrast with kinematic ones.

The machine learning workflow aimed at automatically identifying the signers is displayed in Figure 9.1. The mocap data of the pelvis marker were ignored as it was set as the origin, which leads to zero vectors. Position and velocity of the 18 other markers were used as temporal features. Velocity was estimated by time differentiation of the mocap position coordinates (ORI, SI, SH or POST). Then, we measured time-averaged statistics of these temporal features.

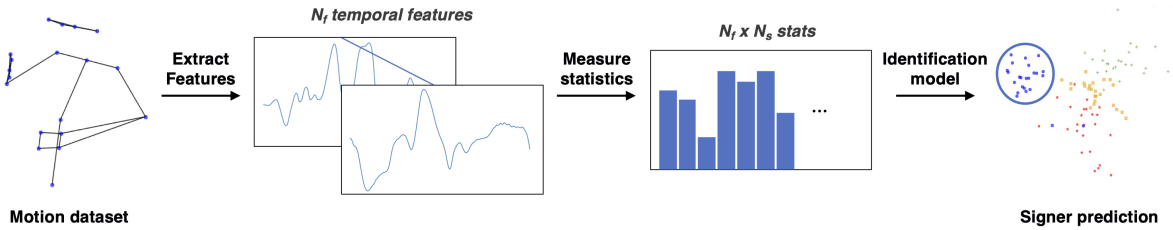


FIGURE 9.1: Schematic representation of the steps used in the machine learning model for identification.

The use of time-averaged statistics, rather than temporal-based methods (for a comparison, see Section 9.2.4), was motivated by the following assumptions. Identity is a time-invariant property that humans are able to recognize from different utterances of the same individual. This makes time-averaged statistics a particularly suited description to extract identity-specific features. In the auditory domain, Latinus and Belin (2011) have shown that speakers' dissimilarities, across brief vowel utterances, were well explained using the average fundamental frequency of phonation (f0) and the average first formant frequency (F1). The role of statistics for categorical discrimination of sounds have further been shown with human behavioral data in McDermott et al. (2013), revealing that discrimination of sounds improved with longer excerpts, notably for the recognition of a single speaker. Converging evidence has also been provided by machine learning of human motion: a linear regression model trained by Tits (2018) has been able to accurately predict the level of expertise from gesture in Taijiquan, based on mean and standard deviation of position and velocity. Moreover, Carlson et al. (2020) recently demonstrated that a dancer's identity may be encoded by the covariance of three-dimensional movements between specific body markers.

Statistics of motion were computed as follows. Based on previous research investigating the perception of auditory and visual textures (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000), we measured the first four moments of position and velocity (Equation 9.1), and covariances of velocity between body markers (Equation 9.2). The first four moments of position and velocity described their statistical distributions, which may vary from one individual to another, as shown for expert gesture analysis (Tits, 2018). For instance, for position, the mean provides information about the average posture of the signers, and the standard deviation provides information about the amplitude of their movements. For velocity, standard deviation provides information about the amount of velocity of a signer's markers in any of the three dimensions. Although the interpretation of the other moments is more challenging, their role in the identification was tested, similarly to McDermott and Simoncelli, 2011. Moreover, the covariance of velocity allowed for quantifying the extent to which any two markers covaried with each other, in two directions.

This latter statistic has been shown to allow for automatic person identification from dance movements (Carlson et al., 2020).

For each mocap example, the triangular part of the covariance matrix was reshaped into a vector of length 1431 and concatenated with the moments of position and velocity, of length 53 each. The concatenated statistics constituted the feature vector used in our person identification model. By definition, posture-normalized data had the same mean position so this latter statistic was not included in POST condition. The computation of the first four moments (Equation 9.1) and covariance (Equation 9.2) is detailed as follows:

$$\begin{aligned} M_{1,m} = \mu_m = \frac{1}{T}\sum_{t=1}^{T} x_m(t), \quad M_{2,m} = \sigma_m = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_m(t)-\mu_m)^2}, \\ M_{3,m} = \frac{\frac{1}{T}\sum_{t=1}^{T}(x_m(t)-\mu_m)^3}{\sigma_m^3}, \quad M_{4,m} = \frac{\frac{1}{T}\sum_{t=1}^{T}(x_m(t)-\mu_m)^4}{\sigma_m^4} - 3 \end{aligned} \tag{9.1}$$

where x_m is the temporal feature (position or velocity) of marker m, along one of the three directions, $m \in [1, 54]$.

$$C_{i,j} = \frac{1}{T-1}\sum_{t=1}^{T}(x_i - \mu_i)(x_j - \mu_j) \tag{9.2}$$

where $x_{i,j}$ are velocity features related to two markers i, j. $\mu_{i,j}$ is the mean of the feature, $i, j \in [1, 54]$.

The relevance of using these statistical measures for signer identification was additionally supported by observations made on our mocap data, which showed that statistics varied substantially across the mocap data of different signers. For instance in Figure 9.2, the position and velocity data of one body marker are distributed differently across signers, for one mocap example (i.e., similar picture description in LSF). Distributions of position data differ in location of the peak (captured by the mean), width (captured by the standard deviation), asymmetry (captured by the skew) and tails created by outliers (captured by the kurtosis).

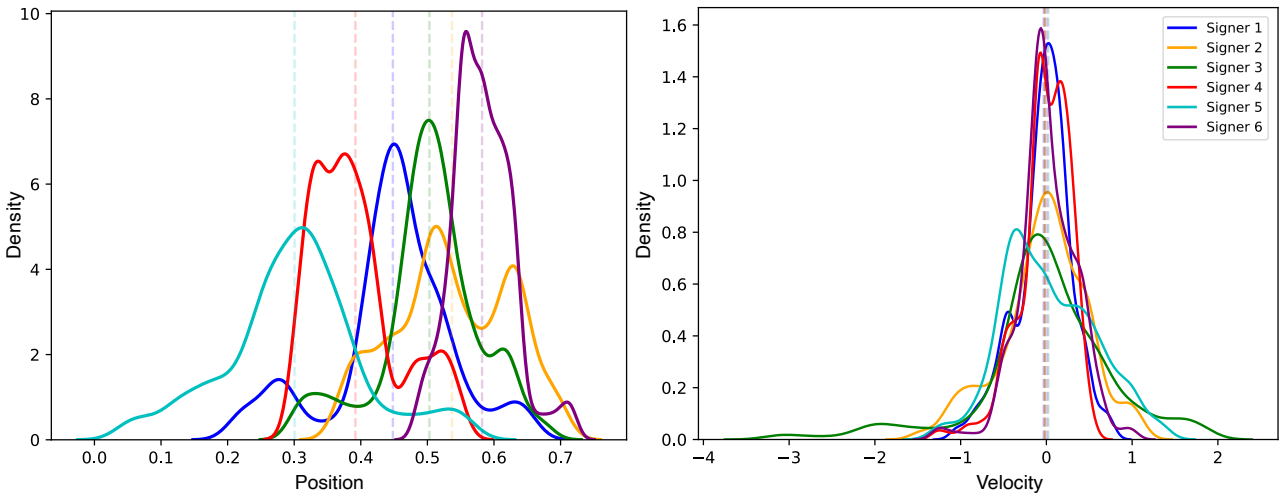


FIGURE 9.2: Distributions of position and velocity data of the RF hand marker along the Z axis, for mocap example 24. Dashed vertical lines represent the means.

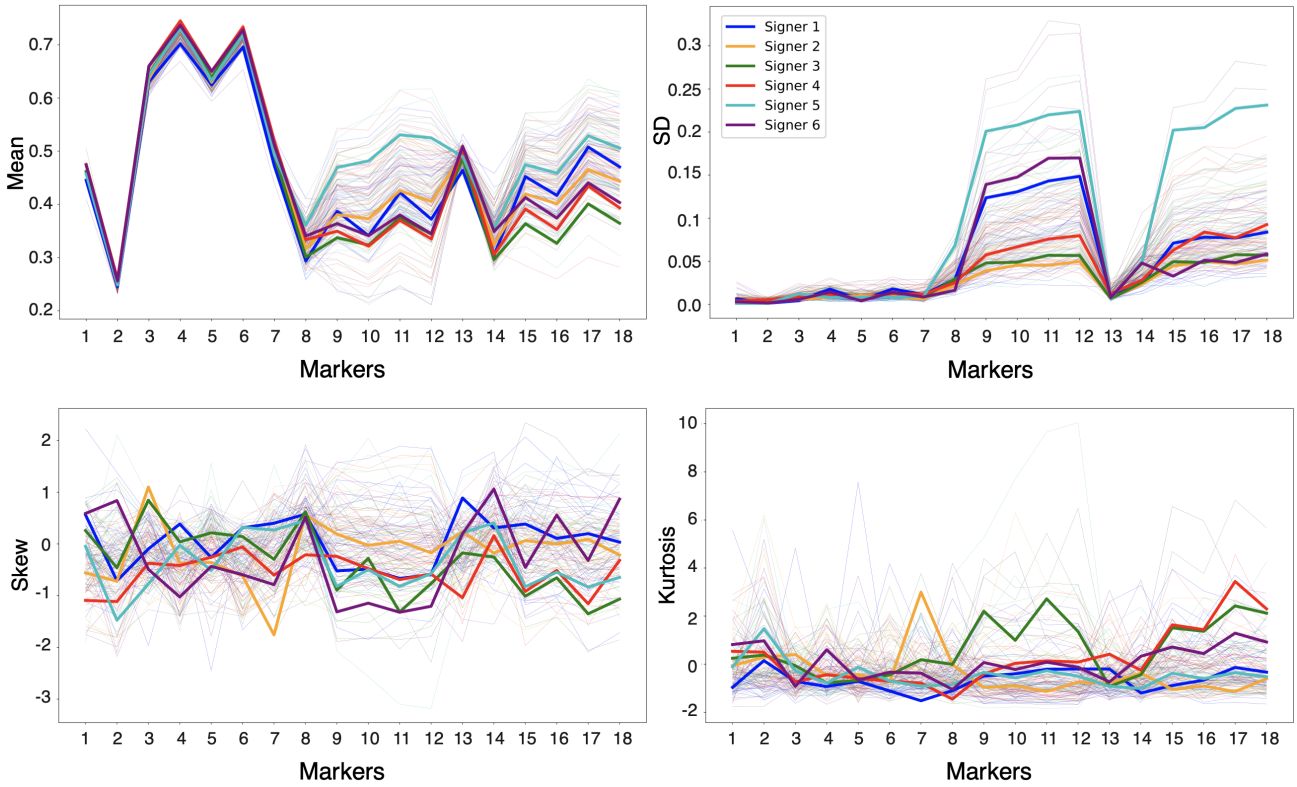


FIGURE 9.3: The four moments of the position data along the Z axis, for all markers and all 144 mocap examples. Thick lines represent the average statistics of each signer across their 24 examples.

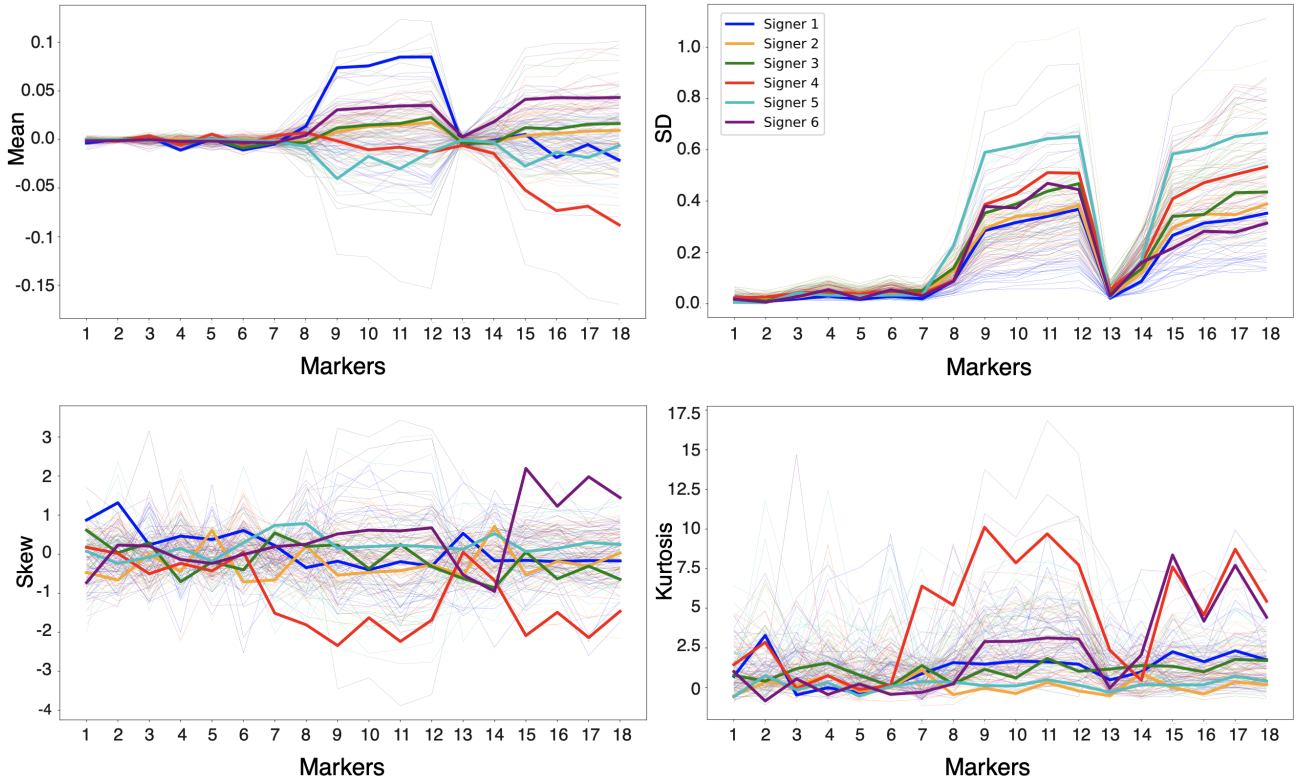


FIGURE 9.4: The four moments of the velocity data along the Z axis, for all markers and all 144 mocap examples. Thick lines represent the average statistics of each signer across their 24 examples.

The extent to which the four moments of position can capture these differences in the distributions was further supported by the values of the moments themselves, as shown in Figure 9.3. Similar observations can be made for the velocity distribution in Figure 9.2, except maybe for the mean, which is similar across signers and near zero for this mocap example. This could be explained by the fact that mocap examples represent continuous SL discourses with posture changes toward both negative and positive directions along the three axes (i.e., making velocity zero-centered) and with breaks (i.e., instantaneous zero-velocity). Still, as shown in Figure 9.4, the four moments of velocity seemed to capture susbtantial differences across signers, mean included. Further tests of the need for all these statistics to correctly identify the signers are presented in Section 9.2.6.

The remaining statistics are the velocity covariances. They capture different aspects of motor coordination between the markers in three dimensions, which can differ across signers. Various distinct coordination patterns can be extracted. For instance, for mocap example 24, the movements of Signer 2 show an overall substantial (positive) covariance between body markers along the Y axis, while this covariance is near zero for Signer 4 (Figure 9.5.A). Inversely, Signer 4 displays an important (negative) covariance of movements of the right arm and hand along the Y axis with the trunk (i.e., stomach and sternum) and head markers along the X axis, while this covariance is less important for Signer 2 (Figure 9.5.B). Various patterns may be extracted from this latter statistic because of the high number of markers and dimensions. To overcome this problem, PCA was used (see 9.1.2), which allowed reducing dimensions of the statistical features and extracting distinct motion patterns that may account for signer identification.

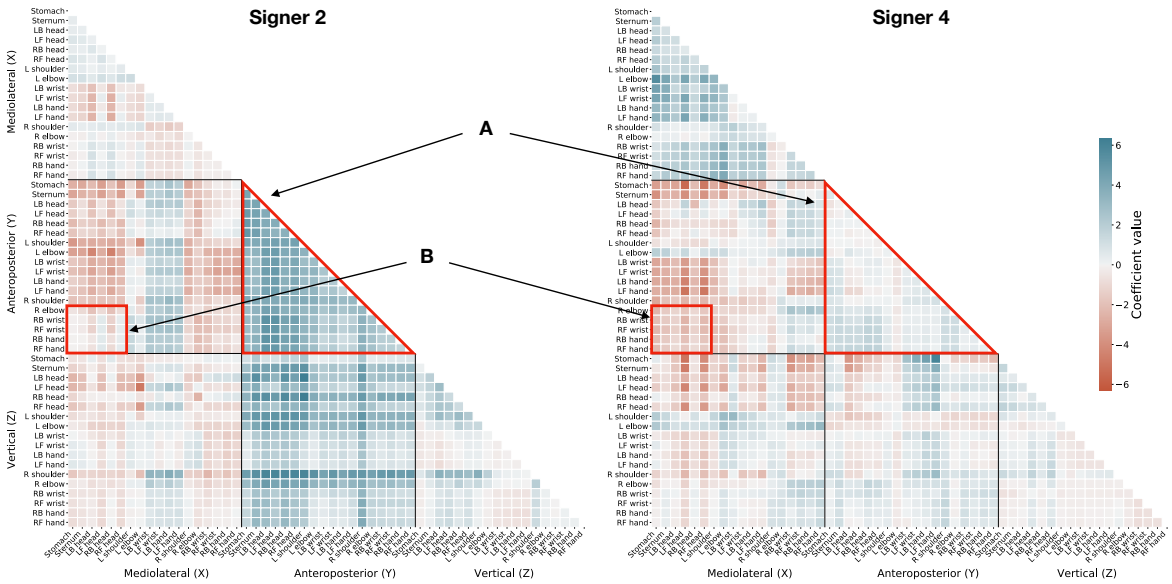


FIGURE 9.5: The covariance of velocity between body markers (rows and columns) of Signer 2 and Signer 4 in the three dimensions, for mocap example 24. Markers are sorted from the 1^{st} to the 19^{th} as presented in Section 5.3.1, along X, Y and Z axes. Coefficients correspond to the covariance measures centered and standardized across examples and signers. Blue represent positive covariances, while red represent negative ones. (A) overall covariance between markers along the Y axis. (B) covariance between the right hand and arm markers along the Y axis, and the trunk and head markers along the X axis.

9.1.2 Person identification model

To predict the identity of the signer, Principal Component Analysis (PCA) followed by a classifier was used. PCA was applied to the motion statistics (contained in a matrix either of length 144 × 1863 for ORI, SI, SH; or 144 × 1809 for POST) and provided uncorrelated Principal Components (PCs) (or eigenvectors), which are linear combinations of the original statistics:

$$\boldsymbol{D} = \boldsymbol{d_0} + \boldsymbol{XV} \tag{9.3}$$

where the matrix $\boldsymbol{D}$ contains the original statistics of all examples, vector $\boldsymbol{d_0}$ contains the average statistics across examples, matrix $\boldsymbol{X}$ contains the coefficients of the original statistics of all examples in the PC space, matrix $\boldsymbol{V}$ contains the PCs (or eigenvectors).

This data-driven method allowed extracting candidate components for the characterization of identity, without a priori hypotheses on the statistics. It also allowed for dimensionality reduction, enabling us to retain a reduced number of PCs. The number of retained PCs has often been chosen on the basis of the amount of variance they explained (Zago et al., 2017a). In the present study, the number of selected PCs was chosen so that it maximized identification accuracy, by testing the model with an increasing number of PCs (based on the descending order of the variance they explained). This follows the approach proposed by O'Toole et al. (1993) who have shown that for face identification, higher-order PCs, which explain only few variance, capture identity-specific features while most of the variance is covered by low-order PCs (see Section 4.2.1).

On the reduced set of PCs, a classifier was trained. We have tested the differences in performance between three important different classifiers: multinomial logistic regression, linear SVM and RBF kernel SVM. The main difference we aimed to test was between linear (i.e., logistic regression and linear SVM) and non-linear models (RBF kernel SVM). Although it is always specific to the dataset used in the study, the optimal classifier choice can be hypothesized based on some machine learning theories (Murphy, 2012). For instance, if N_{feat} (i.e., the number of features) is significantly larger than N_{ex} (i.e., the number of examples), it is general practice to use a linear model, such as logistic regression or linear SVM. If N_{feat} is small and N_{ex} is slightly larger, non-linear models, such as SVM with a kernel, are generally preferred. However, when N_{ex} becomes greatly larger than N_{feat}, non-linear SVMs are hardly optimal, in which case methods for increasing the N_{feat}/N_{ex} ratio should be found and then linear models can be used. In our case, when using all statistics described above, N_{feat} equals 1863 (ORI, SI, SH conditions) or 1809 (POST condition) while N_{ex} equals 144. N_{feat} is thus significantly larger than N_{ex}, which may provide the machine learning model with enough dimensions to allow for a linear separation of the mocap statistics across signers and examples. In that case, a linear classifier could be optimal. To confirm this hypothesis, we assessed the performance of the automatic identification model using the three classifiers (for further details about the machine learning procedure used for automatic identification, see Section 9.1.3). One major interest of SVM classifiers is that they can be applied to non-linear classification problems by using kernels, which allow projecting the original data in a new dimensional space that is linearly separable (Cristianini, Shawe-Taylor, et al., 2000). Still, we additionally tested a linear SVM (i.e., without kernel) as compared with logistic regression, as the two models are built on different methods (i.e., geometrical (SVM) *vs.* statistical (regression) approaches) and may be more optimal

depending on the feature set (e.g., for overfitting problems[1]). The performance of the identification model as a function of the classifier is shown in Table 9.1:

TABLE 9.1: Identification performance of the different classifiers, averaged over the four normalization conditions: ORI, SI, SH and POST.

Classifier	Correct identification
Logistic regression	92.7% (SD = 3.5%)
Linear SVM	91.0% (SD = 4.8%)
RBF kernel SVM	85.8% (SD = 2.7%)

Theoretical predictions were thus confirmed, with an advantage of the linear models over the non-linear one. We further present the classifier that reported the highest performance among linear models, that is multinomial logistic regression. For the prediction of each signer, a logistic regression model was trained, as defined in equation 9.4.

$$P(S = s) = \frac{e^{\boldsymbol{\beta}_s . \boldsymbol{X}}}{\sum_{k=1}^{6} e^{\boldsymbol{\beta}_k . \boldsymbol{X}}} \tag{9.4}$$

where $\boldsymbol{X}$ is the vector containing the coefficients of the test data in the PC space, the vector $\boldsymbol{\beta}_k$ contains the regression coefficients optimized for the identification of Signer k during the learning step, S is the signer variable. The signer s reaching the highest probability P in the model is defined as the predicted signer.

9.1.3 Automatic identification procedure

A leave-one-out cross-validation was conducted: the model was trained on N-1 (23) mocap examples for each signer, and the remaining mocap example was used as the test example (i.e., an unknown example that the model must identify as the signer's production). All examples were used as test example so the model was tested 24 times and performance was computed as an average across these iterations. Using this cross-validation step, we assessed the extent to which the classifier learned idiosyncratic movement statistics that generalize to new mocap examples.

Finally, to better understand the motion statistics that allowed for identification, we scrutinized some discriminant PCs (i.e., PCs that contributed to a significant increase in identification accuracy) in terms of the original statistics they described. First, the general statistical patterns $\boldsymbol{d_n}$ of each PC were described as the absolute value of the PC ($\boldsymbol{V_n}$) (equation 9.5). Based on these descriptions, we then proposed some interpretation of the motion information these PCs might contain. Second, the optimized regression coefficient that the classifier assigned to a given PC for the identification of Signer k was projected onto the PC (equation 9.6). The resulting statistical patterns $\boldsymbol{d_{n,k}}$ provided further insights about the differences between signers along the given PC ($\boldsymbol{V_n}$).

$$\boldsymbol{d_n} = |\boldsymbol{V_n}| \tag{9.5}$$

[1]Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. Depending on the structure of the feature set, logistic regression can be more vulnerable to overfitting problems than SVM.

$$\boldsymbol{d}_{n,k} = \beta_{n,k}\boldsymbol{V}_n \tag{9.6}$$

where $\boldsymbol{V}_n$ is the n^{th} PC (or eigenvector) of the PC space, the scalar $\beta_{n,k}$ is the optimized regression weight assigned to $\boldsymbol{V}_n$ by the classifier to identify signer k. $\boldsymbol{d}_n$ and $\boldsymbol{d}_{n,k}$ are vectors containing the statistical patterns (e.g., of length 1809, in POST condition).

9.2 Results

Size and shape normalizations of the mocap data did not affect the identification performance of the model (Section 9.2.1). Posture normalization caused a significantly lower identification accuracy, but it remained over five times superior to the chance level (Section 9.2.2). The further kinematic cues used by the model to identify the signers from the posture-normalized mocap data defined the kinematic signature of the identity of each signer (Section 9.2.3). The statistics we used allowed for a significantly higher identification performance than temporal descriptors, such as the widely used principal movements (PMs) (Section 9.2.4). Moreover, although the performance of the model was sensitive to the duration of the mocap examples, it remained over two times superior to the chance level when using 0.1-s mocap excerpts (Section 9.2.5). Finally, posterior analyses suggested that some statistical measures (e.g., standard deviation of position and velocity, covariance of velocity) may have a greater importance than others in the automatic identification of signers (Section 9.2.6). This information will be of particular intereset for the development of our further synthesis algorithm, which is aimed at manipulating the discriminant statistics of motion in order to control the identity of signers in SL animations.

9.2.1 The role of structural and kinematic features

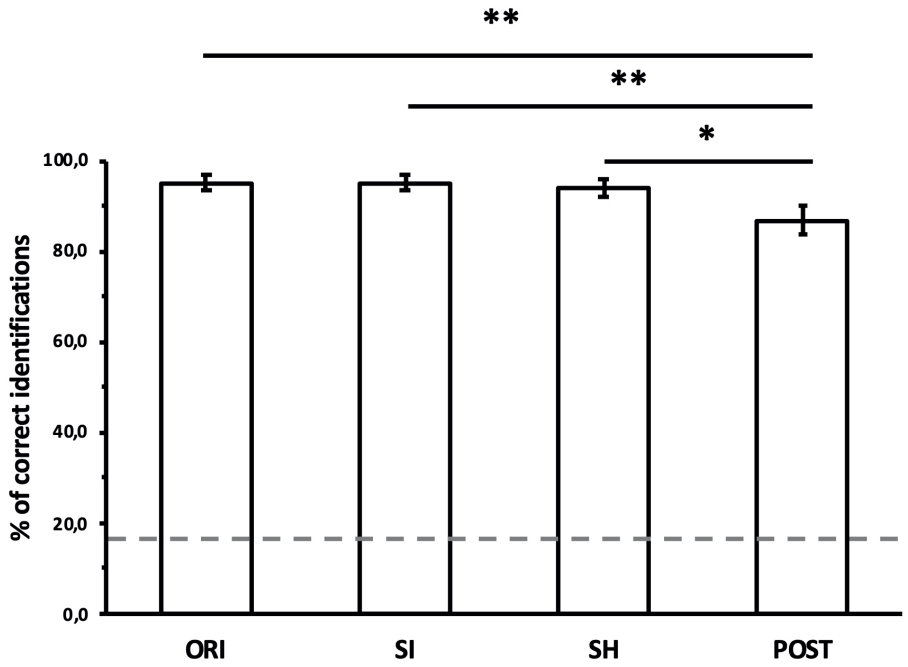


FIGURE 9.6: Average correct identifications of the model, as a function of the normalizations of structural features. ORI: original motion, SI: size-normalized, SH: shape-normalized, POST: posture-normalized. Dashed horizontal line indicates chance level. Error bars indicate standard errors across the 24 test folds. Significant differences between normalizations: *($p < .05$), **($p < .01$).

Correct identifications of the model as a function of the normalizations are shown in Figure 9.6. A repeated measures one-way ANOVA with normalization (with its four levels: ORI, SI, SH and POST) as within-test factor was run on correct identifications. As the assumption of sphericity was violated (Mauchly's Test, $p < .05$), a Greenhouse-Geisser correction was applied ($\varepsilon = .61$). The main effect of normalization was significant ($F(1.83, 42.07) = 5.46, p < .01, \eta^2 = .19$). Bonferroni-adjusted post-hoc tests were performed to test for differences between normalizations. They revealed a significant increase of identification accuracy from posture-normalized (POST, mean = 86.8%) to shape-normalized (SH, mean = 93.8%, $p < .05$), size-normalized (SI, mean = 95.1%, $p < .01$) and original motion (ORI, mean = 95.1%, $p < .01$). No significant difference was found between ORI, SI or SH ($p > .05$).

9.2.2 Identification accuracy of the model for posture-normalized motion

Figure 9.7 displays the correct identifications of the model from posture-normalized mocap data. The number of retained principal components varied from 1 to 144 (which corresponds to the number of mocap examples across signers) (see Section 9.1.2 for details about how the PCs were retained). The highest accuracy of 86.8% was obtained using 69 components. The first component alone allowed for a 38.9% average correct identification. The first 24 components alone contributed to most of the correct identifications, with a 79.2% accuracy. Components 59 to 69 then contributed to most of the increase toward the highest accuracy, from 77.8% to 86.8%.

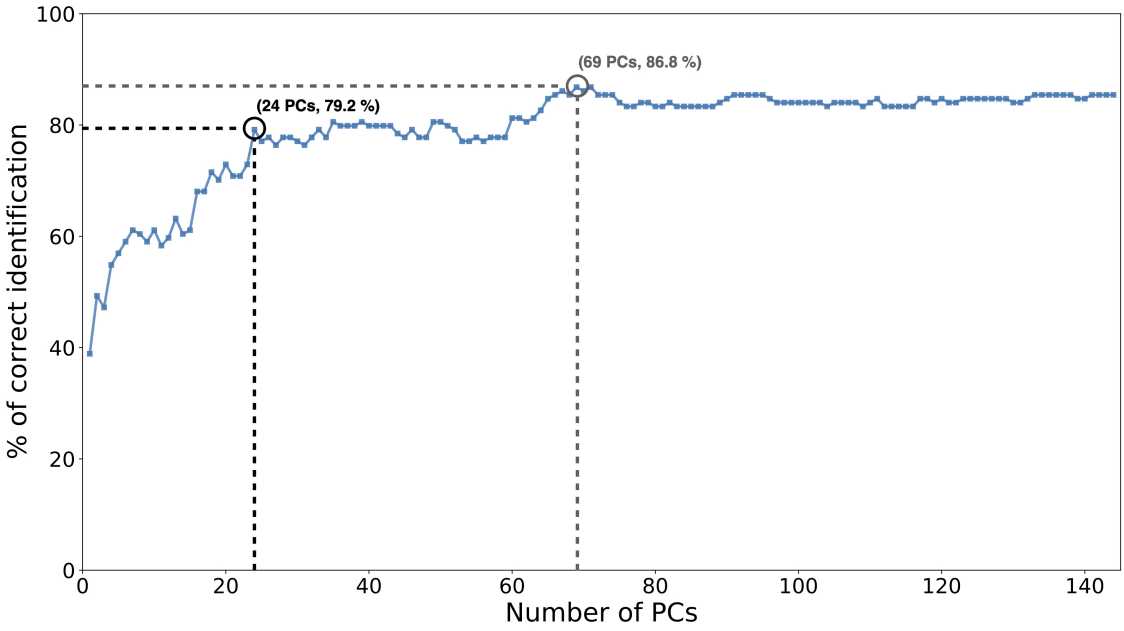


FIGURE 9.7: Correct identifications of the model from posture-normalized motion, as a function of the number of principal components used.

Table 9.2 shows the confusion matrix of the model trained with 69 components, leading to the highest identification accuracy. It specifies the predictions for each signer, across the 24 examples. One sample Student's t-tests revealed that identification performance was above chance level (16.7%) for all signers ($p < .001$ - Signer 1: $t(23) = 9.33, d = 1.91$, Signer 2: $t(23) = 10.27, d = 2.59$, Signer 3: $t(23) = 18.99, d = 4.69$, Signer 4: $t(23) = 7.47, d = 1.53$, Signer 5: $t(23) = 8.58, d = 2.19$, Signer 6: $t(23) = 18.99, d = 4.69$). No confusions were significant between signers ($p > .05$). The lowest performance of the model occurred for Signer 4, with a 70.8% accuracy.

TABLE 9.2: Confusion matrix in percent correct identification of the model, averaged across examples (for posture-normalized motion). Accuracy values significantly above chance level are shown in bold: ***($p < .001$).

	Signer 1	Signer 2	Signer 3	Signer 4	Signer 5	Signer 6
Signer 1	**79.2*** **	0	0	8.3	12.5	0
Signer 2	0	**87.5*** **	4.2	4.2	0	4.2
Signer 3	0	4.2	**95.8*** **	0	0	0
Signer 4	8.3	8.3	4.2	**70.8*** **	4.2	4.2
Signer 5	16.7	0	0	0	**83.3*** **	0
Signer 6	4.2	0	0	0	0	**95.8*** **

9.2.3 Kinematic features of importance

In order to further understand which kind of information is useful for signer identification from posture-normalized motion, we examined the PCs used by the classifier. The identification model was run on the whole dataset with the 69 components, which allowed reaching the highest performance. Discriminant PCs were described following equation 9.5 (see Section 9.1.3). The statistical patterns (referred to as $\boldsymbol{d_n}$ in equation 9.5) of some highly discriminant PCs are displayed in Figure 9.8. PC1, PC2 and PC4 contributed to 38.9%, 10.4% and 7.6% of the cumulative correct identification, respectively (Figure 9.7).

PC1 mainly described relationships between movements along vertical (Z) and anteroposterior (Y) axes, except between hand markers along the Z axis and head markers along the Y axis (Figure 9.8.C). It also described differences in standard deviations of the position and velocity for all body joints along the Y axis (Figure 9.8.A), and for the trunk and head along the Z axis (Figure 9.8.B). PC2 was mostly related to movements along the mediolateral (X) (Figure 9.8.D) and Z axes (Figure 9.8.E). Covarying movements of the head with the right hand along the X axis (Figure 9.8.F) are characteristic of this PC, as well as the right hand with the left hand along Z and X axes, respectively (Figure 9.8.G). PC4 did not describe global movements along one of the three axes, compared with PC1 and PC2. Instead, it mainly characterized relationships between movements along the X and Y axes, particularly regarding the right hand (Figure 9.8.H).

These PCs, either combined or independently, can be used to discriminate between individual signers. For instance, Figure 9.9 displays the idiosyncratic statistical patterns (referred to as $\boldsymbol{d_{n,k}}$ in equation 9.6) of some signers, along PC1. According to PC1, the movements of Signer 1 presented little relationship between anteroposterior and vertical axes (Figure 9.9.C1), and low variation in position and velocity, along anteroposterior and vertical axes (Figure 9.9.A1 and B1). By contrast, Signer 2 was characterized by strong relationship between anteroposterior and vertical axes (Figure 9.9.C2), and high variation in position and velocity, along anteroposterior and vertical axes (Figure 9.9.A2 and B2). These discriminant PCs convey the motion signature of each signer's identity and they can be scrutinized in terms of the original statistics. These findings mean that identity can be inferred from simple statistics of kinematic features, with a consistent accuracy.

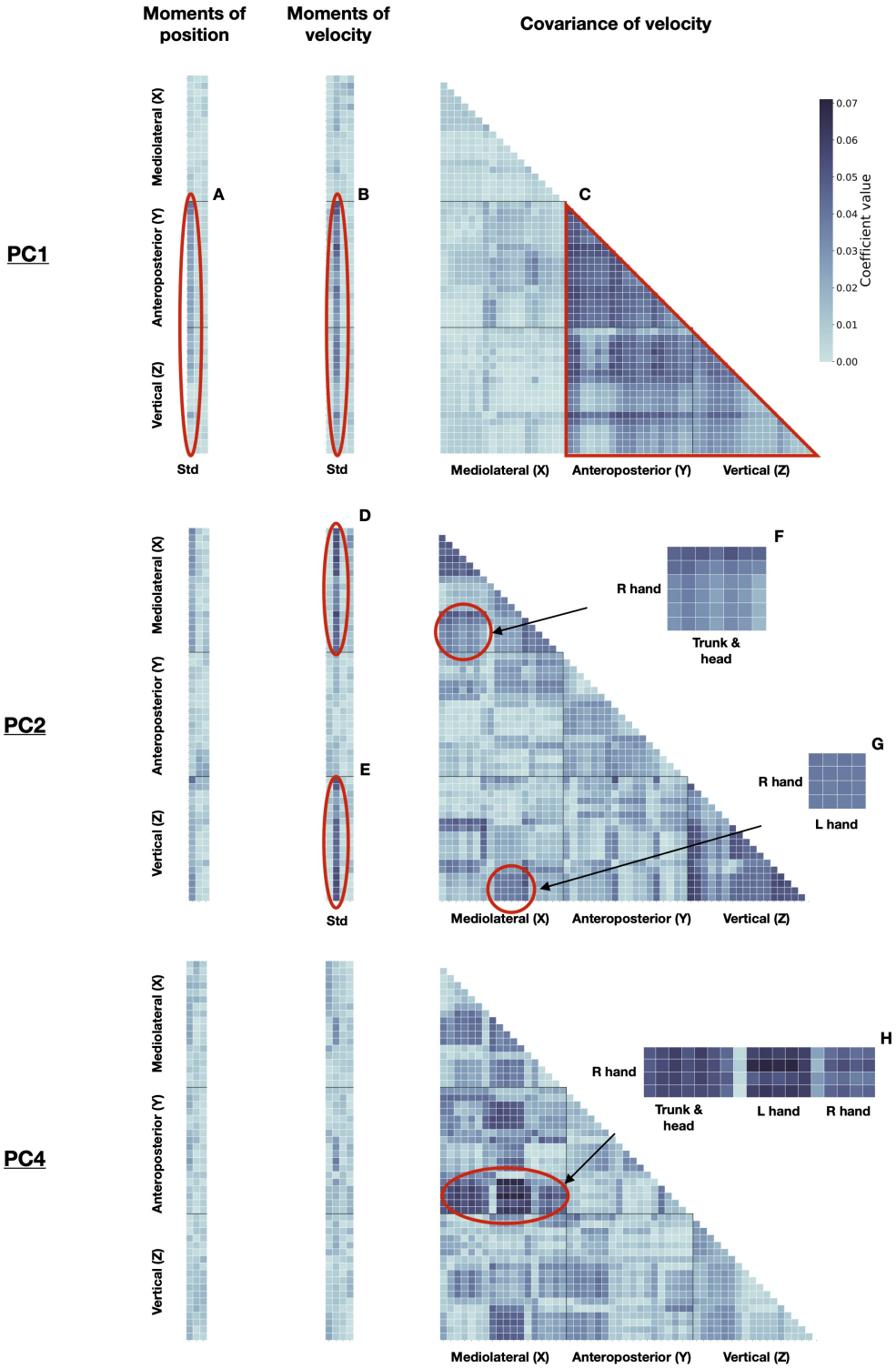


FIGURE 9.8: Discriminant PCs for signer identification. Left: moments (columns: std, skew, kurtosis) of position, for all markers (rows). Middle: moments (columns: mean, std, skew, kurtosis) of velocity, for all markers (rows). Right: covariance of velocity between markers (rows and columns). Markers are sorted from 1 to 19 as presented in Section 5.3.1 along X, Y and Z axes. Some patterns of importance are highlighted. For sake of clarity, the specific moments and body markers are displayed only for these patterns of importance. PC1: Std of position (A) and velocity (B) along Y and Z axes, for all markers; (C) Covarying movements between all markers along Y and Z axes. PC2: Std of velocity along X (D) and Z (E) axes, for all markers; (F) Covarying movements between the right hand, and trunk and head markers, along X axis; (G) Covarying movements between the right hand markers along Z axis, and the left hand markers along X axis. PC4: (H) Covarying movements between the right hand markers along Y axis, and all other markers along X axis.

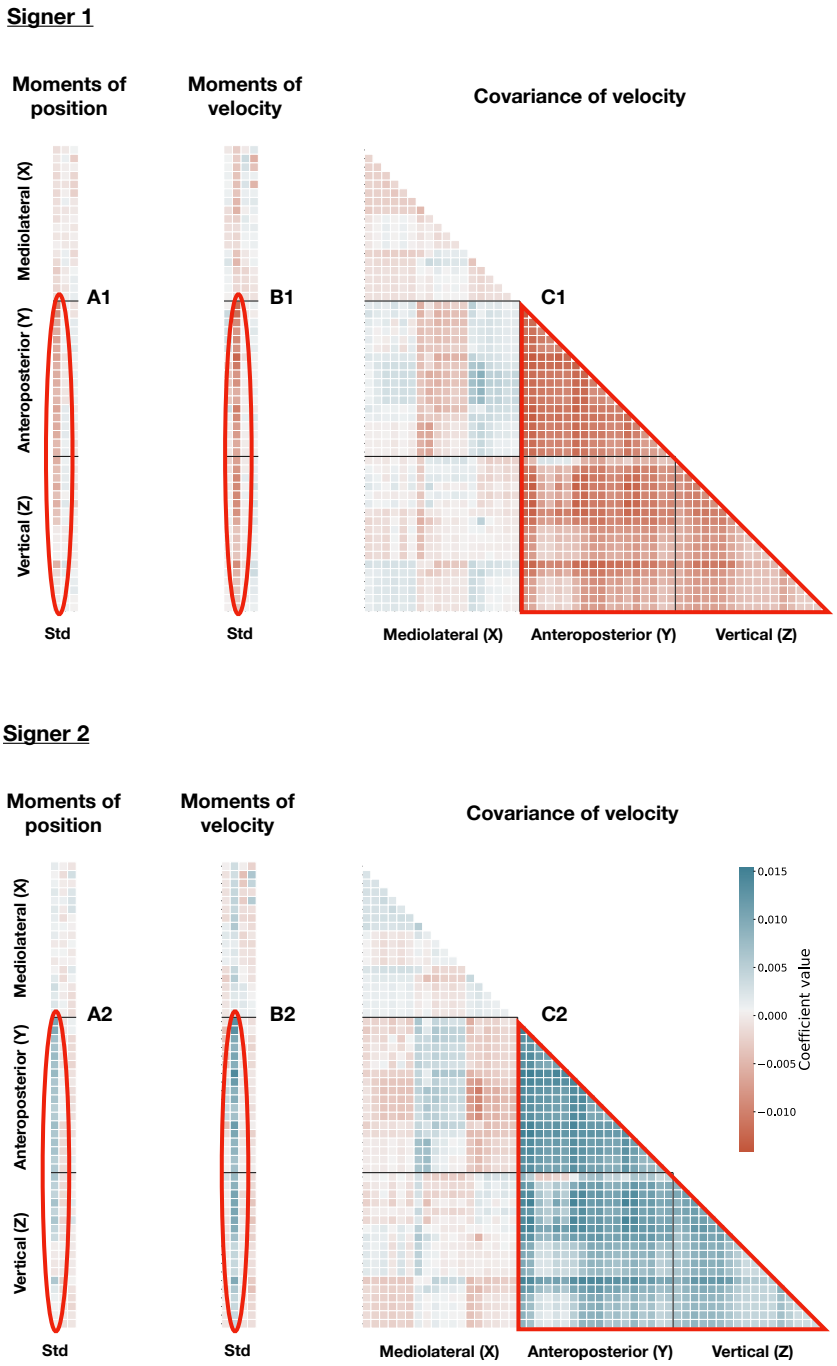


FIGURE 9.9: Classifier weights of PC1 for Signer 1 and Signer 2. Similarly to Figure 9.8, for Signer 1 (left) and Signer 2 (right): moments (columns: std, skew, kurtosis) of position for all markers (rows), moments (columns: mean, std, skew, kurtosis) of velocity for all markers (rows), covariance of velocity between markers (rows and columns). Markers are sorted from 1 to 19 as presented in Section 5.3.1 along X, Y and Z axes. Coefficients correspond to the logistic regression weights optimized for each signer. Blue represents positive weight values, while red represents negative ones. The three patterns of importance are highlighted: Std of position (Ak) and velocity (Bk) of all body markers for Signer k; (Ck) Covariance of velocity between all body markers along Y and Z axes for Signer k.

9.2.4 The overall advantage of statistics over temporal-based approaches

As mentioned in Section 4.2.3 previously and in Section 9.1.2 of the present chapter, we hypothesized that statistics of the mocap data would be particularly suited for signer identification, as compared to temporal-based descriptions. The results of the present study then confirmed that a statistical-based model could allow for successful automatic signer identificaiton (Section 9.2.2) and determination of the specific motion features used in the identification (Sections 9.2.1 and 9.2.3). To further assess the relevance of choosing a statistical-based approach, we tested our machine learning model on a widely used temporal-based representation: principal movements (PMs). In Young and Reinkensmeyer (2014), the quality of dives was automatically evaluated across various athletes, using PM posture vectors and their temporal weights, jointly with more traditional features used for dive evaluation (e.g., splash area). PM posture vectors of individual walkers have also allowed for automatic gender classification (Troje, 2002a) and person identification (using key postures similar to PMs although they were obtained from Fourier decomposition) (Zhang and Troje, 2005).

In the present study, we trained our model to identify signers from the temporal weights of the PMs, similarly to Tits (2018) and Zago et al. (2017a). Although more complete feature representations allowed for even more accurate prediction, Zago et al. (2017a) have shown that PM temporal weights could allow predicting the expertise of karateka. Moreover, as demonstrated in Chapter 7, signers of MOCAP1-v2 shared common PMs but executed them differently. Eight common PMs were sufficient to explain 95% of variance in the original LSF movements of all of the six signers. Therefore, the model was trained on the weights of these eight common PMs. In other words, we assessed the extent to which the differences in the execution of the common PMs between signers allowed identifying them with high accuracy, compared to our statistical-based approach. The automatic identification model (see Section 9.1.2) and procedure (Section 9.1.3) used with statistics and principal movements were identical. Finally, the identification performance of the model was compared between the two approaches. As shown in Figure 9.10, the highest identification accuracy using PM weights is 31% and is obtained with 27 PCs. Our statistical-based approach outperforms the one based on PM weights, with a 86.8% accuracy obtained with 69 PCs.

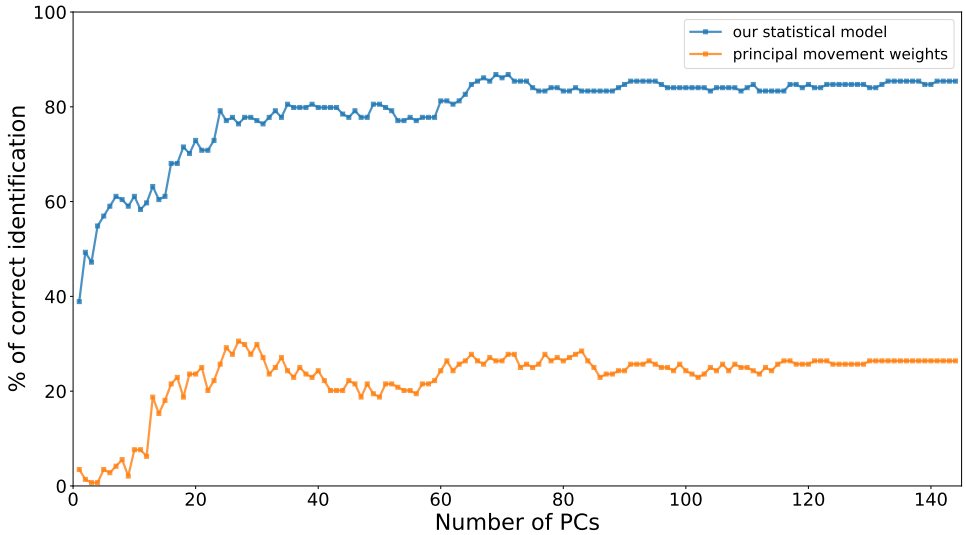


FIGURE 9.10: Identification performance of our model using a statistical-based approach, compared with our model using a temporal-based approach (i.e., principal movement weights). Performance is plotted as a function of the number of PCs used in the PCA step of the machine learning framework (see Section 9.1).

9.2.5 Fast identification of signers: to which extent?

The results mentioned above pointed out an overall advantage of statistical over temporal representations for the identifiaction of signers from our mocap dataset, using 5-second mocap examples. Would statistics also allow for successful identification in very short periods of time? Indeed, statistics are averaged over time, which makes them sensitive to the duration of the mocap original examples. Beyond the fact that one may expect identification to be easier with motion examples of longer duration, we were interested in the extent to which very fast examples could still allow for substantial identification. Humans are very fast (< 0.2 seconds) at recognizing movement categories from PLDs (Johansson, 1976), as in the auditory and vision domain, where human perceivers can categorize sounds (Bigand et al., 2011; Agus et al., 2012) and human faces (Rousselet et al., 2003) very rapidly (< 500 milliseconds). To address that question for the identification of signers from motion, we trained our statistical-based model using mocap examples of various durations. As expected, the identification accuracy of the model improved as a function of the duration of the mocap excerpts (Figure 9.11), from 49.3% with 0.1-second excerpts to 86.8% with 5-second ones. When using examples of duration longer than 2 seconds, the model managed to identify the signers with accuracy rates above 80%.

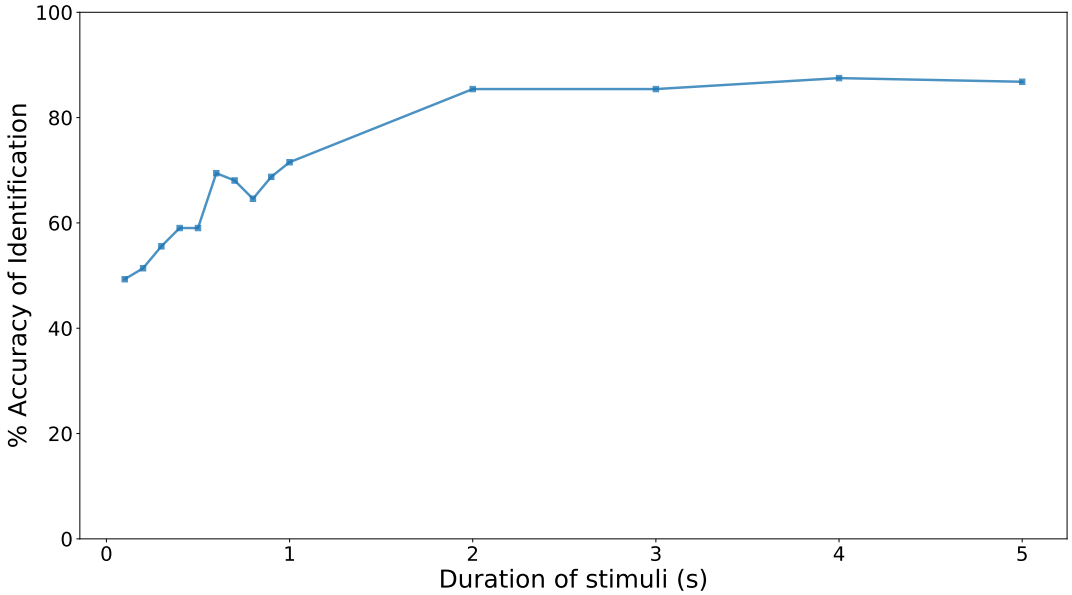


FIGURE 9.11: Correct identifications of our model as a function of the duration of the mocap examples.

Still, the fact that very fast, 0.1-second, mocap examples allowed for a substantial (i.e., 49.3%) identification accuracy of the model was quite intriguing. We hypothesized that our trimming method (i.e., the start frame of all mocap examples was set just after the initial "T" posture) may have influenced the model and allowed for such performance, assuming that signers could have made a similar movement when beginning all their discourses. Therefore, a second analysis was then conducted where the start frame was set to a random time (within a range of four seconds) after the end of the initial "T" posture, rather than 0 initially. For instance, the start frame could be set to 2.3 seconds after the "T" posture for one example, while being set to 3.8 seconds for another example of the same signer, which reduced the potential influence of a "typical" movement made by the signers at the

beginning of all their discourses. All mocap examples were still of 5-second duration. As shown in Figure 9.12, the identification accuracy of the model was lowered for rapid excerpts (e.g., from 49.3% to 38.2% for 0.1-second mocap excerpts) when the start frames were set randomly. More interestingly, the gap in identification performance between random and non-random conditions reduced as the duration of the stimuli increased. If the performance of the model was 11.1% lower for 0.1-second excerpts in the random condition than in the non-random one, it was only 2.8% lower for 5-second ones. These results are consistent with our hypothesis that initial movements of the signers used to begin their discourses may have had a slight influence on the predictions of the model. Indeed, these movements were potentially important in short excerpts, while their importance may be reduced when averaging the statistics on longer durations, for the benefit of other movements specific to the discourse. Still, the 38.2% accuracy reported for 0.1-second mocap examples in the random condition remains substantial, compared with the chance level (i.e., 16.7%). This suggests that identity could be inferred from the movements of signers very rapidly and emphasizes the idea that signer identification may be achieved beyond semantic content, which is unlikely to be comprehensible in such short times.

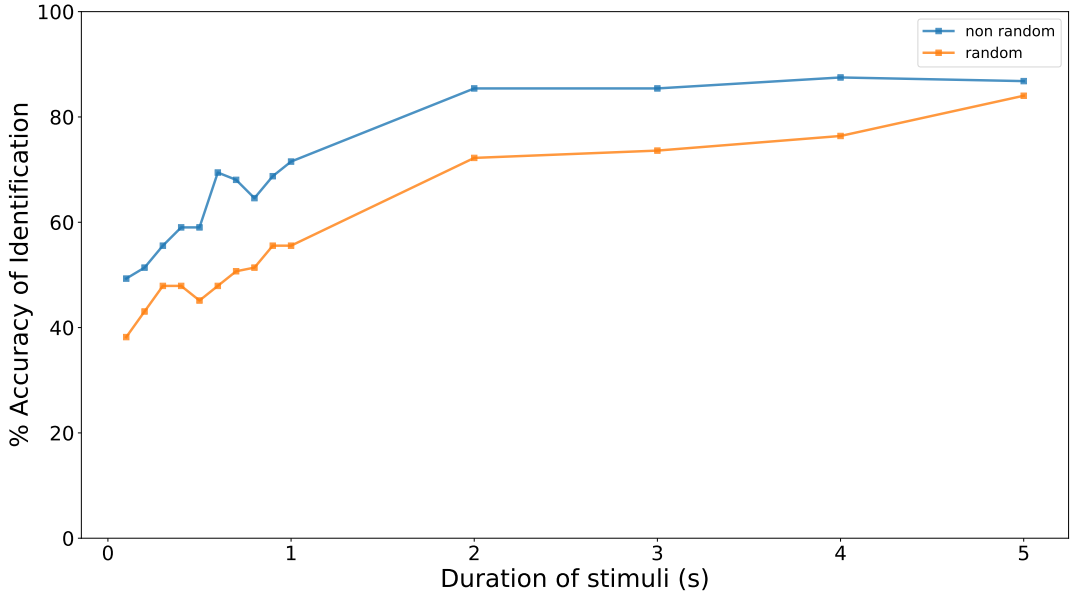


FIGURE 9.12: Correct identifications of our model as a function of the duration of the mocap examples, in the two conditions: non-random (i.e., all mocap examples are trimmed just after the end of the initial "T" posture) (blue) and random (i.e., the start frame after the end of the "T" posture is set to a random value, between 0 and 4 seconds, for each mocap example) (orange).

9.2.6 The statistics: all needed?

Finally, we assessed the necessity of using all statistical measures in our identification model. As pointed out in Section 9.1.1, some statistics used in our model could potentially contribute less to the identification than others. For instance, Figure 9.2 showed that the mean values of the velocity data were quite similar across signers. We thus evaluated the identification accuracy of the model from posture-normalized motion using different subsets of statistics. Interestingly, the model managed to identify the signers with substantial accuracy (i.e., 65.3%, approximately four times higher than the chance level) using the standard deviation (SD) of position

data. The main significant increase in accuracy between conditions occurred when adding the SD of velocity data ($p < .05$). Overall, performance of the model significantly increased from the input matrix containing only the SD of position (65.3%) to the whole matrix containing all statistics (86.8%, $p < .001$). Extensive statistical differences between conditions (i.e., using the different subsets of statistics) can be found in Appendix C. The best identification performance was thus achieved using all statistics. However, the three statistics that seemed to play a major role in the ability of the model to identify the signers were SD of position, SD of velocity and covariance of velocity. These results form the basis of the statistics that will be manipulated in our further synthesis algorithm aimed at controlling identity-specific features in SL movements (see Chapter 10).

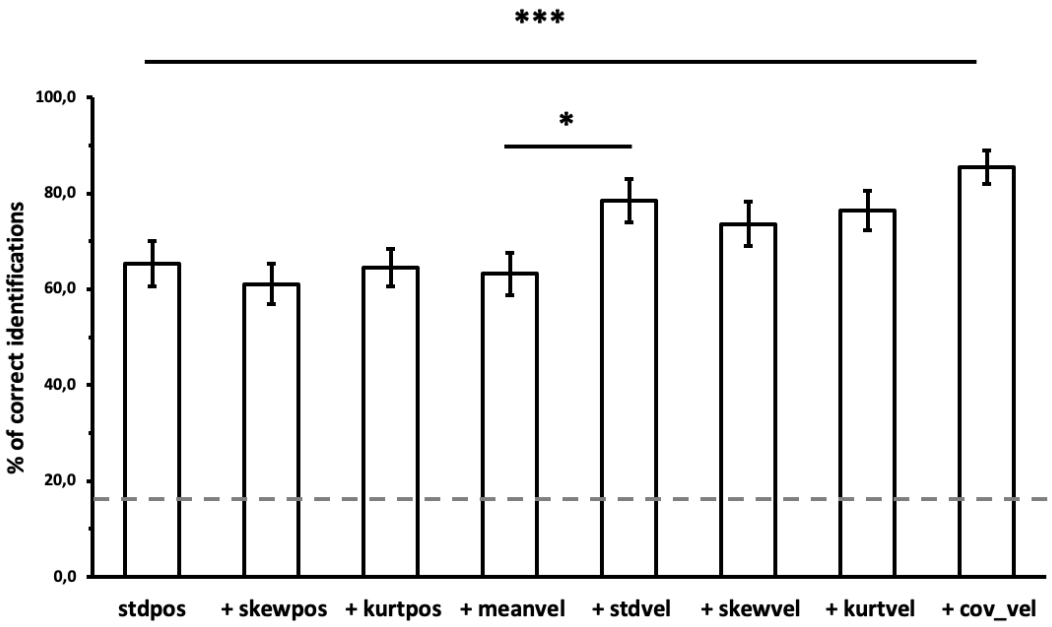


FIGURE 9.13: Average correct identifications of the model from posture-normalized motion, as more statistics are used as input features. Two important significant differences are shown between conditions (for the further differences between all conditions, please refer to Appendix C). Dashed horizontal line indicates chance level. Error bars indicate standard errors across the 24 test folds. Significant differences between conditions: *($p < .05$), ***($p < .001$).

9.3 Discussion

The present study demonstrates that motion capture data convey critical information to allow for robust identification of signers using machine learning, as previously shown for walking (Zhang and Troje, 2005) or dancing (Carlson et al., 2020). PCA followed by a linear classifier managed to correctly identify signers from the statistics of their movements recorded during the free description of pictures in spontaneous French Sign Language (LSF). Even when deprived of structural information about the signers, the model reported 86.8% accuracy, over five times higher than chance level (Sections 9.2.1 and 9.2.2). These results are consistent with prior findings on the human ability to identify individuals from walking (Cutting and Kozlowski, 1977; Troje et al., 2005; Westhoff and Troje, 2007) and dancing (Loula et al.,

2005; Bläsing and Sauzet, 2018) movements. In particular, this is in line with the outcomes of Chapter 8, which demonstrated that humans are able to identify signers from PLDs of their movements in LSF. Compared to the latter visual perception experiment, which measured the human ability to identify signers, the present study trained a machine learning model, which successfully identified signers from statistics of mocap data. Although in Chapter 8, we have shown that size and shape of the signers' body may not have played a major role in the ability of the participants to identify, the machine learning approach taken in the present chapter allowed further determining the specific features that allow for the identification.

The second outcome of the present study is that kinematics alone allow for robust identification of the signers. Removing size and shape information did not affect the performance of the model. Normalizing the mocap data with respect to the signers' postures led to a decline of identification accuracy. Nevertheless, the remaining identification accuracy was significantly above chance level. The minor role of anthropometric differences in identifying individuals from their movements is consistent with prior behavioral studies on gait (Troje et al., 2005; Westhoff and Troje, 2007) and LSF (see Chapter 8). Interestingly, the impact of the signers' average posture on our model's correct identification was similar to the impact reported by Troje et al. (2005) on human observers, causing a decrease of about 10% of accuracy. The remaining ability of the model to identify signers without any of these structural cues confirms that kinematics alone are sufficient to achieve identification, as previously suggested by Troje et al. (2005) and Westhoff and Troje (2007) for walking.

Further analyses of the contribution of kinematic features to the model's identification revealed identity-specific characteristics of signers' motion (Section 9.2.3). In general, discriminant PCs described specific kinematic statistics in all three dimensions. For instance, PC1, PC2 and PC4, which accounted for 54.9% of correct identification, were characterized by movements in the sagittal, frontal and transverse plane, respectively. Previous findings have outlined critical features for gender classification of gait in the frontal plane, which are, therefore, best visible in frontal view (Mather and Murdoch, 1994; Troje, 2002a). However, although Troje et al. (2005) have found an overall advantage for walker identification based on the frontal view, training on half-profile views allowed for higher performance when participants had to identify walkers from new viewpoints. Moreover, no overall advantage for the frontal view has been reported by Westhoff and Troje (2007), whose gait PLDs were totally deprived of structural information. Whereas for now, human perceivers' ability to identify signers have only been studied using frontal views (Chapter 8), half-profile and profile views may provide critical information, especially for kinematics. This observation is consistent with the recent machine learning model of dancer identification proposed by Carlson et al. (2020), which reported kinematic features of importance along all three dimensions.

Similarly to Carlson et al. (2020), the discriminant PCs revealed distinct identity-specific patterns over sensors and dimensions. For instance, whereas PC1 reflected differences in the kinematics of all body markers along the anteroposterior axis, differences along the vertical axis concerned only the trunk (e.g., stomach, sternum, shoulders) and head markers. PC1, PC2 and PC4 reported different contributions of each part of the signers' bodies, often distinguishing groups of markers such as head, trunk or hand markers. We also noticed distinct contributions of the two hands, such as a lower impact of the left hand along the mediolateral axis in PC2 than the right hand. This may be due to the motion differences caused by the dominant hand of the signers, which was the right hand for all of them. Indeed, as with any other human movements, signers preferably use their dominant hand when signing, such as

for pointing, fingerspelling, asymmetric two-handed signs (i.e., the dominant hand moves with respect to the other hand, which stays still) or one-handed signs (i.e., the sign is executed with the dominant hand only), as this hand provides faster or more precise performance. Prior studies have highlighted inter-individual differences in the execution of principal movements (or eigenmovements) for skiing (Federolf et al., 2014), karate (Zago et al., 2017a) or pathological gait (Zago et al., 2017c). However, principal movements are based on frame-by-frame relations between gestures, while SL movements are hardly ever synchronized across examples and individuals. Hence, as previously pointed out by Tits (2018), we outlined here the advantage of using statistics as motion descriptors for identity, which is invariant to time and independent of semantic content, over temporal-based approaches.

Unlike prior work on the recognition of familiar faces (O'Toole et al., 1993), which had demonstrated that identity may be better characterized by high-order PCs of the face space (i.e., with low eigenvalue) (see Section 4.2.1), most of the PCs that carried critical information for the identification of signers in our study were in lower dimensions. Indeed, the first 24 PCs allowed the model to identify the signers from posture-normalized motion, with a 79.2% accuracy. Beyond the differences in the data representations to which PCA was applied (i.e., pixel gray levels of face images in O'Toole et al. (1993) *vs.* statistics of mocap data in our study), further differences in the structure of our dataset can have caused this low-dimensional representation of identity in the movements. In O'Toole et al. (1993), 159 different faces were used (i.e., 159 examples for 159 identity labels), while we used 24 mocap examples per signer, resulting in 144 examples across all of the six signers (i.e., 144 examples for six identity labels). We could hypothesize that the PCA interpreted information about the identity of the six signers as category information that is shared by multiple examples (of the same signer) and thus represented in lower PC dimensions, rather than as identity-specific information that is not shared by examples (i.e., related to one specific example, as in O'Toole et al. (1993)) and thus better characterized by higher PC dimensions. In other words, our machine learning model may have interpreted the six identity labels of the signers as six categories across mocap examples (like gender was represented as two categories across face examples in O'Toole et al. (1993)). This observation calls for further research investigating signer identification from mocap datasets with more signers, each represented by one mocap example only.

The differences we found in the information carried by low- and high-order PCs was related to the statistical measures. For instance, PCs 1 to 24 were mainly related to the standard deviation of position, standard deviation of velocity and covariance of velocity. By contrast, PCs 59 to 69 mainly corresponded to finer patterns of skew and kurtosis measures. This is in line with further analyses we conducted on the specific importance of each statistical measure in the automatic identification (Section 9.2.6). Although it was not clear whether the contribution of skew and kurtosis measures could be considered as negligible, the statistics of greater importance in the identification were the standard deviation of position, standard deviation of velocity and covariance of velocity. Moreover, the mean velocity did not seem to play a major role in the identification, as anticipated in Section 9.1.1. Additionally, the impact of the duration of mocap examples on the performance of the model was assessed (Section 9.2.5). The results showed that signer identification was still possible to some extent with very short examples. For instance, signers were identified by the model with a 38.2% accuracy (i.e., twice the chance level) when using 0.1-second mocap excerpts. These results are in line with the ability of human observers to process

motion patterns from PLDs in very short times, such as 0.2 seconds for recognizing human actions or 0.1 seconds for perceiving a human body in the movements (Johansson, 1976).

The results of the present study suggest that signers have a kinematic signature, which is invariant to the semantic content of their movements in LSF. We were able to characterize this signature using 24 components extracted from PCA, leading to a 79.2% identification accuracy. Such a data-driven approach is particularly interesting in the case of identification as the discriminant features are mainly idiosyncratic and thus hard to define a priori for each individual. The other main advantage of PCA is its invertibility, which makes it possible to recompute statistics by projecting a linear combination of PCs back into the original space. These statistics could be manipulated in order to resynthesize pre-recorded SL movements isolating, or exaggerating, the PCs of interest. To achieve this, we could develop algorithms similar to the ones used to synthesize sounds with matching statistics (McDermott et al., 2009; McDermott and Simoncelli, 2011; Norman-Haignere and McDermott, 2018). This approach would allow for the visualization of specific PCs, by exaggerating their weight in the combination of PCs, as previously shown for male and female gaits (Troje, 2002a). Furthermore, being able to control identity-specific PCs in motion synthesis would provide promising perspectives toward anonymizing SL motion for virtual signers.

Chapter 10

Synthesis algorithm for the kinematic control of identity

The primary motivation of the present thesis was that novel engineering tools were needed to allow controlling the features that carry identity information in the movements of virtual signers, in particular for generating anonymized content (see Chapter 1). Compared to speaker identification in spoken languages, little was known about the critical motion features that allowed for signer identification in Sign Languages. From both human and computational data (see Chapters 8 and 9), this thesis has demonstrated that cues related to structural differences may not play a major role in the identification of signers, except ones related to their posture, which may provide a partial account for the identification. Identity may thus be mainly inferred from the kinematic aspects of the movements. Using the machine learning model developed in Chapter 9, we are now able to automatically extract the specific kinematic aspects of motion that carry identity using time-averaged statistics. Manipulating these discriminant statistics in the generation of SL movements could allow changing the identity perceived by the human observers (e.g., the movements of the signer could be anonymized). This final contribution of the thesis presents a synthesis algorithm developed in order to manipulate identity-specific kinematic statistics from original mocap recordings (Section 10.1). Performance of the algorithm is assessed in terms of convergence and quality of statistical matching, and is illustrated with some examples of kinematic anonymization and identity conversion using SL mocap data of MOCAP1-v2 (Section 10.2). Although this version of the algorithm is a first prototype, it provided convincing results and opens up promising perspectives toward the automatic control of identity in the movements of virtual signers, in the same way as for the voice of a speaker, which can be anonymized by modifying specific vocal parameters (Section 10.3).

10.1 Methods

The automatic signer identification model presented in Chapter 9 allowed extracting specific kinematic statistics that carry identity information about the signers. An original synthesis algorithm was further developed, which allowed reducing or exaggerating these statistics in novel, synthesized, mocap examples in order to change the identity inferred from the movements of the signers (Section 10.1.1). Moreover, as previously outlined in this thesis (see Sections 1.2 and 3.3.3), one crucial challenge of automatic SL generation is to keep the movements of virtual signers natural and comprehensible. Therefore, not only was our synthesis algorithm aimed at imposing new statistics but it also needed to preserve the original structure of the movements, notably to keep the semantic content intact (Section 10.1.2).

10.1.1 Imposing new statistics to the movements

The synthesis was driven by the discriminant statistics extracted by the signer identification model of Chapter 9. As detailed in equation 9.6, the identity of each signer k was characterized by discriminant statistical patterns $\boldsymbol{d_{n,k}}$, each related to a specific PC $\boldsymbol{V_n}$. Using the 24 PCs that allowed for a 79.2% identification accuracy of the model (see Chapter 9), the overall discriminant statistical pattern of the identity of each signer was formulated as follows:

$$\boldsymbol{d_k} = \sum_{n=1}^{24} \boldsymbol{d_{n,k}} \tag{10.1}$$

where $\boldsymbol{d_{n,k}}$ is a vector containing the statistical patterns of the n^{th} discriminant PC for signer identification and $\boldsymbol{d_k}$ is a vector containing the overall statistical patterns of the first 24 discriminant PCs for Signer k.

The aim of the present synthesis algorithm was to impose new target statistics $\tilde{\boldsymbol{d}}_{\boldsymbol{\alpha}}$ to an original mocap recording of a given signer in order to reduce ($\alpha < 0$) or exaggerate ($\alpha > 0$) the identity-specific aspects of motion that are characteristic of Signer k, following Equation 10.2:

$$\tilde{\boldsymbol{d}}_{\boldsymbol{\alpha}} = \boldsymbol{d_{orig}} + \alpha \boldsymbol{d_k} \tag{10.2}$$

where $\tilde{\boldsymbol{d}}_{\boldsymbol{\alpha}}$ is a vector containing the new target statistics imposed by the synthesis alogrithm, $\boldsymbol{d_{orig}}$ is a vector containing the original statistics of the mocap example, $\boldsymbol{d_k}$ is a vector containing the overall statistical patterns related to the identity of Signer k, and α is a scalar related to the amount of reduction ($\alpha < 0$) or exaggeration ($\alpha > 0$) of the identity attribute.

The different steps of the synthesis process are displayed in Figure 10.1. In summary, the synthesis process consisted of modifying (i.e., "re-synthezing") an existing mocap recording in order to change the identity attribute of the signer, according to the following steps. First, statistics of the original mocap example are measured (for further descriptions of the processing of mocap recordings and calculation of the statistics, see Section 9.1.1), while the discriminant statistical kinematic patterns are taken from the automatic identification model (see Chapter 9). Then, the discriminant statistics characteristic of Signer k are either added to ($\alpha > 0$) or subtracted from ($\alpha < 0$) the ones of the original example (see Equation 10.2). Multiple manipulations can then be done using this technique, depending on the values of k and α. For instance, if the original mocap example relates to Signer 1, reducing the importance of her identity-specific statistics (i.e., $k = 1, \alpha < 0$) would make her less identifiable (i.e., kinematic anonymization). By contrast, increasing the importance of the identity-specific statistics of Signer 2 (i.e., $k = 2, \alpha > 0$) would make this latter signer identifiable while the SL movements were originally executed by Signer 1 (i.e., kinematic identity conversion). Once the target statistics defined, they are imposed to the original mocap signal by the algorithm, which creates a new mocap excerpt.

Target statistics were imposed using an iterative process where a synthesized mocap signal (initialized with the content of the original mocap recording) is modified until its statistics are sufficiently close to the target ones $\tilde{\boldsymbol{d}}_{\boldsymbol{\alpha}}$. Mathematically, the objective of this process is to minimize the loss function that calculates the mean square of the differences between the target statistics and the statistics of the synthesized movements (see Equation 10.3). In the prototype presented in this chapter, we imposed the first two moments (mean and standard deviation (SD)) of position and velocity data, and the covariance of velocity between markers. SD of position, SD of

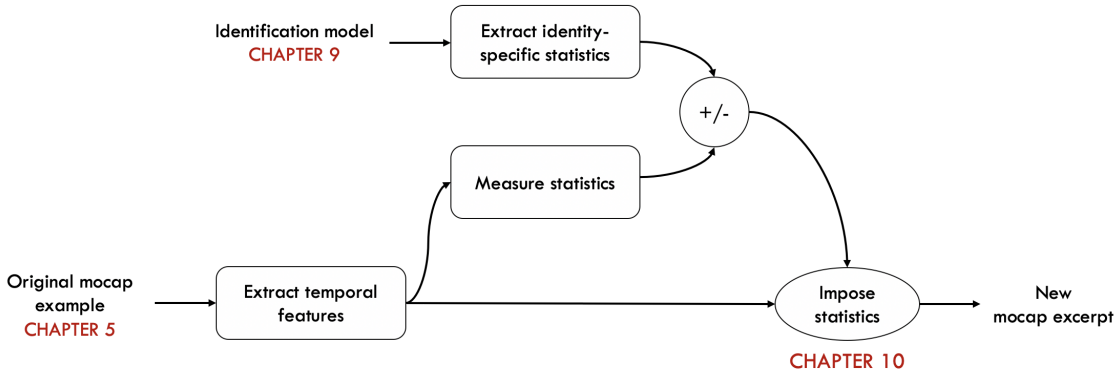


FIGURE 10.1: Schematic representation of the steps used in the synthesis algorithm for the kinematic control of identity.

velocity and covariance of velocity were shown to be the most important statistics that allowed for signer identification (see Section 9.2.6). Imposing the mean of position and mean of velocity of the markers was done to maintain consistent motion data when synthesizing (e.g., to avoid the generation of unrealistic, non-biological, movements), although these two statistics had only minor role in the identification.

$$\begin{aligned} loss = &\sum_{m}(\mu_{pos,m,\boldsymbol{target}} - \mu_{pos,m,\boldsymbol{synth}})^2 + \sum_{m}(\sigma_{pos,m,\boldsymbol{target}} - \sigma_{pos,m,\boldsymbol{synth}})^2 \\ &+ \sum_{m}(\mu_{vel,m,\boldsymbol{target}} - \mu_{vel,m,\boldsymbol{synth}})^2 + \sum_{m}(\sigma_{vel,m,\boldsymbol{target}} - \sigma_{vel,m,\boldsymbol{synth}})^2 \\ &+ \sum_{i,j}(C_{i,j,\boldsymbol{target}} - C_{i,j,\boldsymbol{synth}})^2 \\ &= loss_1 + loss_2 + loss_3 + loss_4 + loss_5 \end{aligned} \tag{10.3}$$

where $\mu_{pos,m}$, $\sigma_{pos,m}$, $\mu_{vel,m}$ and $\sigma_{vel,m}$ are the first two moments of position and velocity data of marker m ($m \in [1,54]$), $C_{i,j}$ is the covariance of velocity between markers i and j. ***target*** and ***synth*** subscripts distinguish between target statistics and statistics of the synthesized movements, respectively.

In order to be able to minimize all of the five loss components of Equation 10.3 despite the differences in ranges of amplitude across statistics, we used a weighted loss function, whose weights then need to be optimized (see Equation 10.4). The loss function was then minimized using the Adam optimization algorithm for gradient descent. Each iterative step of the gradient descent modified the synthesized mocap signals (i.e., position temporal curves of the 19 markers along the three dimensions) so that they approached the target statistics.

$$loss = w_1 loss_1 + w_2 loss_2 + w_3 loss_3 + w_4 loss_4 + w_5 loss_5 \tag{10.4}$$

10.1.2 Preserving the original motion structure

Initially, there was no constraint in the synthesis process that forced the position and velocity signals of the synthesized movements to remain consistent with their initial temporal structure in the original movements. The limitation of this first version of the algorithm is that, although it managed to impose the statistics present in Equation 10.3, the modifications applied to the new movements seemed to generate noise artifacts rather than changing relevant aspects of the motion of the signer (see Video

10.1). This problem is particularly visible in the temporal curves of position and velocity data, as shown in Figure 10.2. In fact, the imposing algorithm managed to impose the target statistics but by modifying the movements in an undesired manner. First, low-energy segments of the motion were modified in the same way as high-energy ones, which is not relevant as they may not be perceived by observers. Moreover, reaching the target statistics caused very rapid oscillations in the synthesized velocity temporal curves, which are unlikely to be perceived as biological motion by the observers (but rather noisy, wobbling markers).

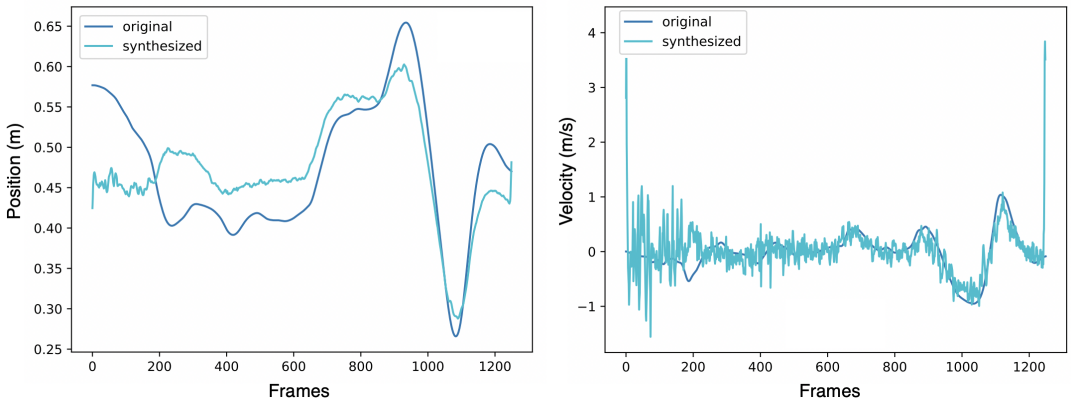


FIGURE 10.2: Example of the synthesis results with the first algorithm version, for mocap example 1 of Signer 1. Position (left) and velocity (right) data of RF hand marker along the Z axis are shown, for the original mocap recording and synthesized mocap excerpt.

In order to modify the movements in proportion to their energy (i.e., modify the aspects of the movement at relevant times of actual, perceptible, motion), we included another target statistic in the imposing algorithm: the correlation of velocity between the original and synthesized movements. The algorithm then aimed to minimize the mean squared error between this correlation and a value of 1, which characterizes two signals that are perfectly positively correlated (see Equation 10.5). In other words, imposing this additional statistic (Equation 10.6) allowed forcing the velocity curves of the synthesized movements to be consistent with their initial temporal structure in the original mocap recording (Figure 10.3).

$$loss_6 = \sum_m (\rho_{vel,m,\boldsymbol{target}} - \rho_{vel,m,\boldsymbol{synth}})^2 = \sum_m (1 - \rho_{vel,m,\boldsymbol{synth}})^2 \tag{10.5}$$

$$loss = w_1 loss_1 + w_2 loss_2 + w_3 loss_3 + w_4 loss_4 + w_5 loss_5 + w_6 loss_6 \tag{10.6}$$

where $\rho_{vel,m}$ is the correlation of velocity between the original and synthesized movements of marker m ($m \in [1, 54]$). The target correlation value is set to 1 for all markers, in order to preserve the original temporal structure of velocity curves.

As shown in Figure 10.3 (and Video 10.2), the second version of the imposing algorithm managed to reach the target statistics but in a more relevant manner than in its first version. Velocity data is highly more realistic, in particular as fast noisy oscillations are removed. Moreover, reaching the target statistics modified the movements at more relevant moments. For instance, the SD of velocity increased from the original to the synthesized movement by generating wider movements (e.g., frames

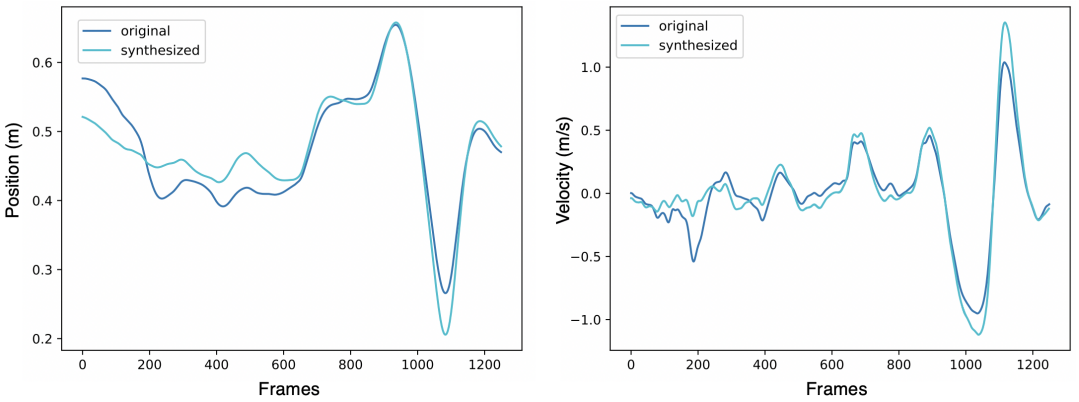


FIGURE 10.3: Example of the synthesis results when additionally imposing the correlation between velocity curves of original and synthesized mocap examples, for mocap example 1 of Signer 1. Position (left) and velocity (right) data of RF hand marker along the Z axis are shown, for the original mocap recording and synthesized mocap excerpt.

1000 to 1200 in Figure 10.3 (left)) with velocity peaks of greater importance (e.g., frames 1000 to 1200 in Figure 10.3 (right)). More global modifications of the marker position also occurred (e.g., frames 0 to 600 in Figure 10.3 (left)) in order to reach all of the target statistics, as compared to the first algorithm version, which caused unrealistic modifications of the position at the frame level (e.g., frames 0 to 200, and 1200 to the end, in Figure 10.2 (left)).

In summary, the imposing algorithm iteratively modified the original movements until their statistics (i.e., first two moments of position and velocity, and covariance of velocity between markers) approached the target ones defined by the user to modify the identity attribute. Moreover, the velocity curves of the original and synthesized movements were forced to be correlated in order to maintain a consistent temporal structure. In the version of the imposing algorithm presented in this thesis, parameters (see Table 10.1) were optimized manually for each synthesized motion example.

TABLE 10.1: Summary characteristics of the imposing algorithm.

Statistics imposed	$\mu_{pos,m}$, $\sigma_{pos,m}$, $\mu_{vel,m}$, $\sigma_{vel,m}$, $C_{i,j}$ and $\rho_{vel,m}$ (see Equations 10.3 and 10.5)
Imposing method	Minimizing the loss function (see Equation 10.6) using gradient descent (Adam optimizer)
Parameters	– Weights of the loss function – Number of iterations of the gradient descent optimizer – Step size of the gradient descent optimizer

10.2 Results

Using a sufficient number of iterative steps, the imposing algorithm managed to modify the movements so that their statistics approached the target ones (Section 10.2.1). This synthesis procedure was run on mocap examples of different signers and for different modifications of the identity attribute. For instance, the movements of Signer 1 were modified so that the perceived identity was that of Signer 2 (i.e., identity conversion) (Section 10.2.2). Then, they were modified to make Signer 1 not

identifiable, without making another signer identifiable specifically (i.e., anonymization) (Section 10.2.3). One further synthesis example of identity conversion (from Signer 2 to Signer 1) is available in Appendix D.

10.2.1 Algorithm validation: convergence and statistical matching

In this section, the performance of the synthesis procedure is illustrated by the first synthesized example presented in Section 10.2.2: identity conversion from Signer 1 to Signer 2, for mocap example 1. For this example, the synthesis was driven by the following target statistics:

$$\tilde{\boldsymbol{d}}_{\mathbf{100}} = \boldsymbol{d}_{orig} + 100\boldsymbol{d}_{\mathbf{2}} \tag{10.7}$$

where $\boldsymbol{d}_{orig}$ are the statistics of the mocap example 1 of Signer 1, and $\boldsymbol{d}_{\mathbf{2}}$ are the discriminant statistical patterns of Signer 2.

The gradient descent procedure was run with 20,000 iterations and a step size of 0.001. The Adam optimizer converged to a low, but non-zero, value. In other words, statistics of the synthesized movements may have approached the target ones but not perfectly, as shown in Figures 10.5, 10.6 and 10.7.

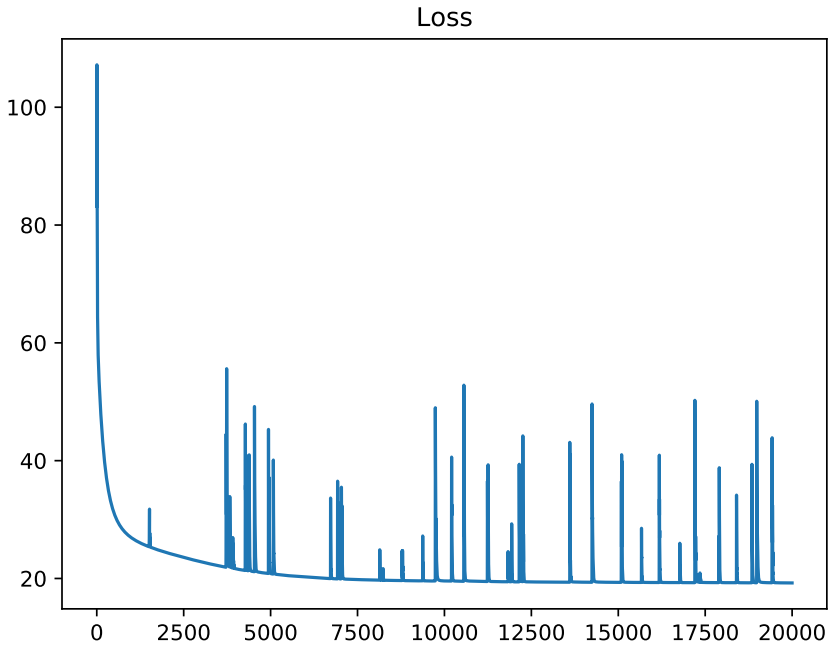


FIGURE 10.4: Loss curve for the kinematic identity conversion from Signer 1 to Signer 2, using mocap example 1 of Signer 1.

As shown in Figure 10.5, target statistics differ from the ones of the initial synthesized movements to various extents across markers. These gaps between target and synthesized statistics depend on the discriminant identity-specific pattern that is manipulated $\boldsymbol{d}_k$ (in this example, $k = 2$) and on the amount of modification α that is set. For instance, modifying the identity attribute in the movements toward Signer 2 involves significantly increasing the SD of position of the right hand markers along the X axis (see Figure 10.5). By contrast, it involves reducing the SD of position of the left hand markers along the Z axis. Then, the value of α defines the extent to which all of these gaps are amplified ($\alpha > 0$) or reduced ($\alpha < 0$). As shown in Figure 10.5, although a perfect maching of the statistics is not reached, the synthesis procedure

allows the SDs of position across markers of the synthesized motion to approach the target. For instance, the significant gap between the SD of position of the right hand markers before the synthesis is now filled perfectly (or almost, for the RF hand marker).

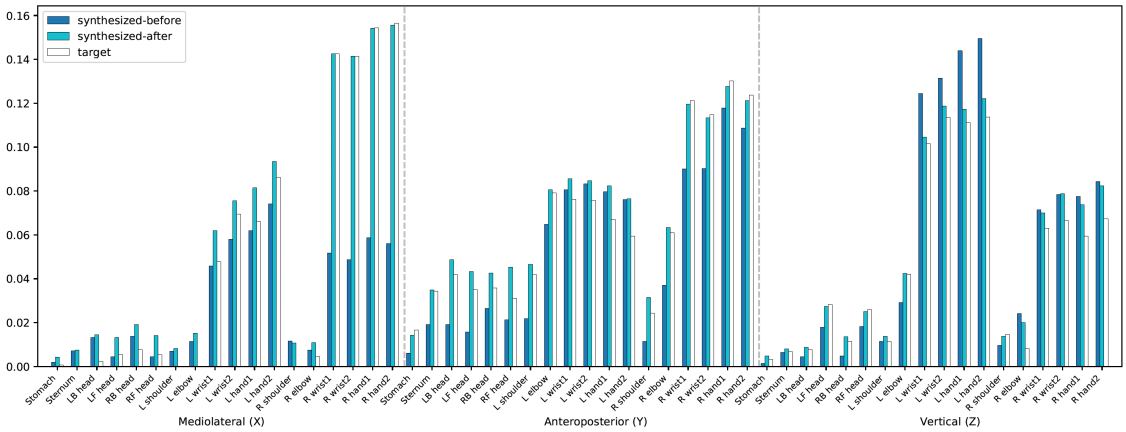


FIGURE 10.5: Standard deviation of position for all body markers before and after the synthesis procedure, for mocap example 1 of Signer 1. Statistics of the synthesized movements are shown in dark blue (before the synthesis procedure) and in cyan (after the synthesis procedure). Target statistics are shown in white. Markers are sorted from the 1^{st} to the 19^{th} along X, Y and Z axes.

Similar observations were made for the SD of velocity, as shown in Figure 10.6. Main increases toward the target values can be seen for both hands along the X axis, for the left hand along the Y axis and for the left hand along the Z axis (see Figure 10.6). Some decreases toward the target values were involved for the right hand markers along the Y axis. As for the SD of position, our algorithm allowed the synthesized movements to approach the target statistics. Statistics along the X axis were almost perfectly matched (see Figure 10.6), while some approximations occurred along the Y and Z axes, which notably caused higher values than the desired ones.

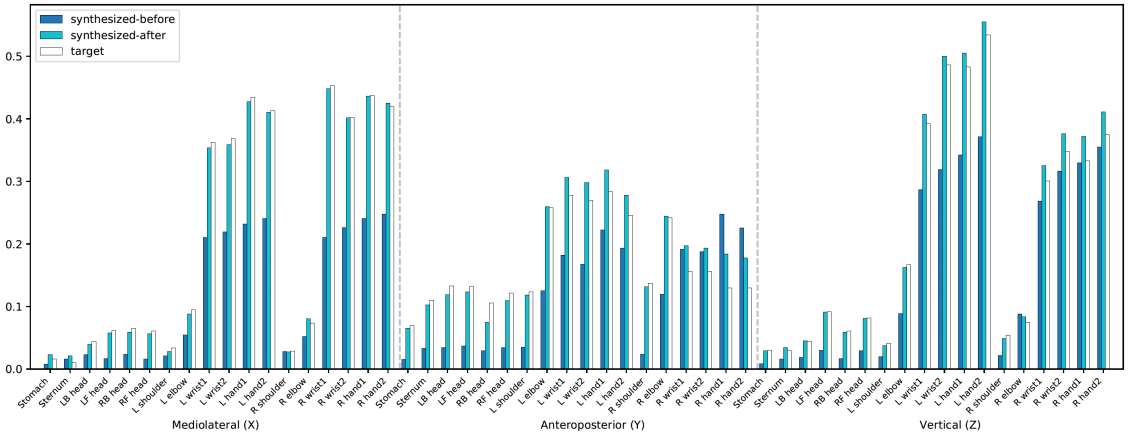


FIGURE 10.6: Standard deviation of velocity for all body markers before and after the synthesis procedure, for mocap example 1 of Signer 1. Statistics of the synthesized movements are shown in dark blue (before the synthesis procedure) and in cyan (after the synthesis procedure). Target statistics are shown in white. Markers are sorted from the 1^{st} to the 19^{th} along X, Y and Z axes.

The quality of matching of the statistics between the target and the synthesized movements was weaker for the covariance of velocity. As shown in Figure 10.7, although some covariance values of the synthesized motion approached their target values, others remained virtually unchanged. Yet, most of the discriminant statistics approached their target values, which should have an effect on the perceived identity of the signers in their movements. In order to assess the extent to which the novel movements generated by our algorithm could convey a modified identity attribute (e.g., could be anonymized, or identified as movements of another signer), we tested our automatic signer identification model (see Chapter 9) on the synthesized mocap examples. If the identity-specific aspects of the movements are correctly modified by the synthesis algorithm, then automatic identification from these synthesized examples should be compromised.

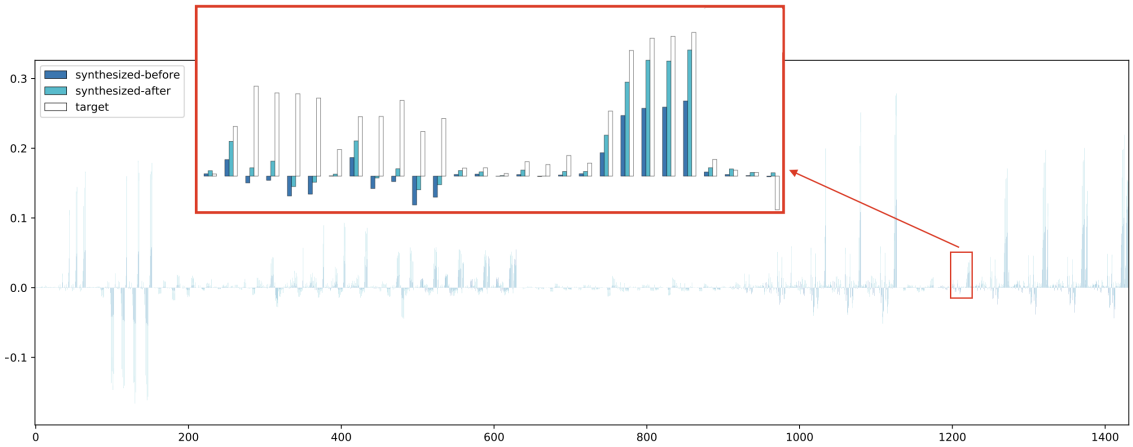


FIGURE 10.7: Covariance of velocity for all body markers and along the three dimensions before and after the synthesis procedure, for mocap example 1 of Signer 1. Statistics of the synthesized movements are shown in dark blue (before the synthesis procedure) and in cyan (after the synthesis procedure). Target statistics are shown in white. Covariance values are sorted from markers covarying along the X axis to those covarying along the Z axis. The figure is zoomed in on some covariance values in order to illustrate the degree of matching between the target and synthesized statistics.

10.2.2 Example 1: identity conversion from Signer 1 to Signer 2

In this first example, the synthesis of the mocap excerpts was driven by the statistics of Equation 10.7. The synthesis procedure involved 20,000 iterations and a step size of 0.001, as explained in Section 10.2.1. Important statistics that carry identity information, such as SD of position, SD of velocity and covariance of velocity between markers, were modified through the process (see Figure 10.8). For instance, there was an overall increase of the SD of velocity for all markers along the Z axis. The covariance of velocity between markers was further interestingly modified. For instance, the synthesis procedure forced the movements of the left hand along the X axis to significantly covary (negatively) with the trunk and head markers along the Y and Z axes. These covarying movements of the trunk and head with the left hand may be characteristic of Signer 2.

In order to visualize how these new statistics affected the movements of the SL discourse of Signer 1, the original and synthesized mocap examples can be seen as Point-Light Display (PLD) videos (see Videos 10.3 and 10.4). Additionally, Figure

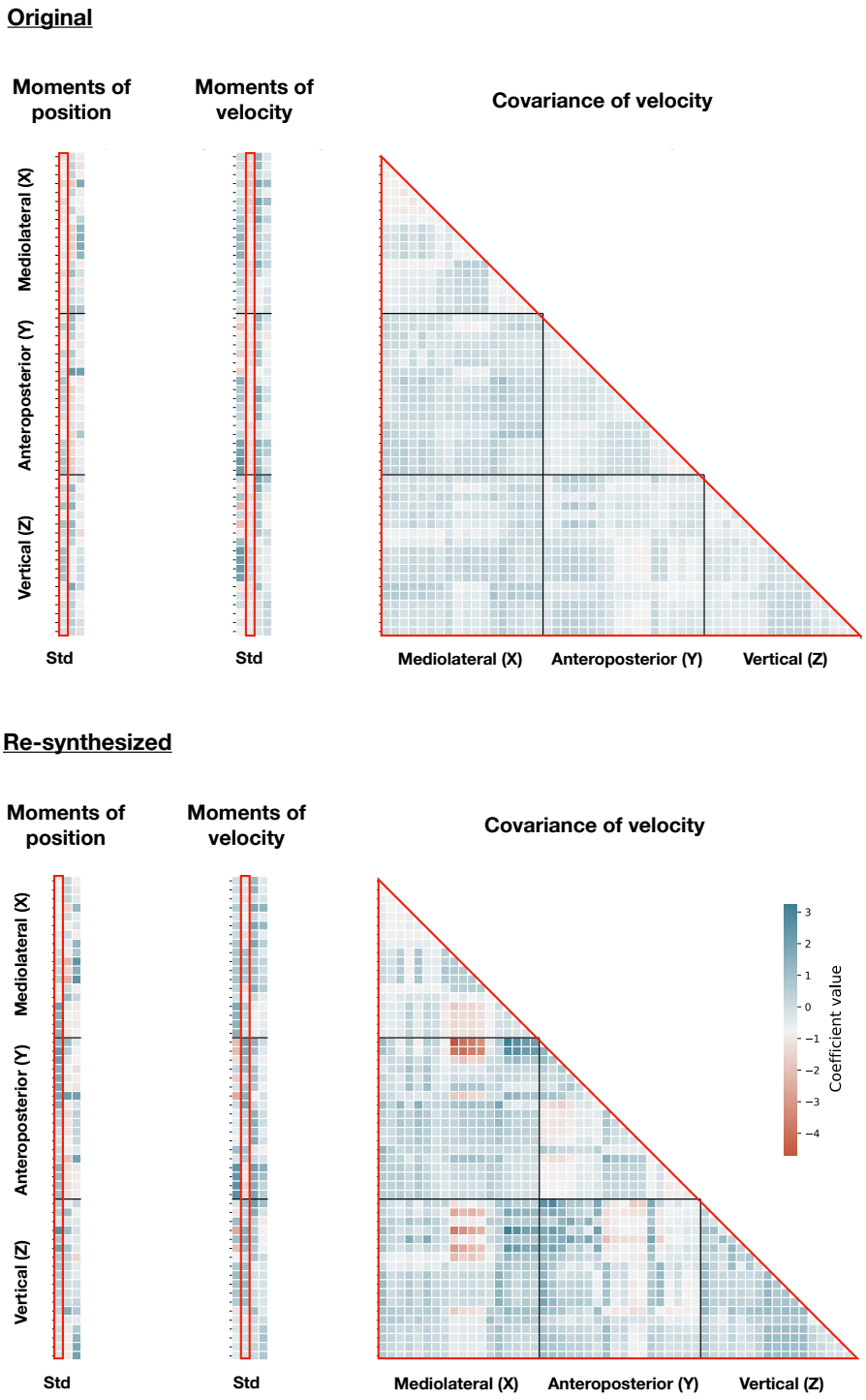


FIGURE 10.8: Statistics of kinematic features of mocap example 1 of Signer 1, before (original) and after (re-synthesized) the synthesis procedure: moments (columns: std, skew, kurtosis) of position for all markers (rows), moments (columns: mean, std, skew, kurtosis) of velocity for all markers (rows), covariance of velocity between markers (rows and columns). Markers are sorted from 1 to 19 as presented in Section 5.3.1 along X, Y and Z axes. Blue represents positive statistical values, while red represents negative ones. The three main classes of statistics that carry identity information are highlighted.

10.9 illustrates how the position and velocity of the RF hand marker were changed. In line with the observations made in Section 10.2.1 from Figure 10.5, the SD of position of this marker were reduced along both X and Y axes. By contrast, the SD of velocity was increased along the X axis while reduced along the Y one, as expected from Figure 10.6. The automatic signer identification model identified the synthesized mocap example as that of Signer 2, as shown in Table 10.2:

TABLE 10.2: Output of the automatic signer identification model. P is the probability that the movements were produced by the signer (see Equation 9.4).

Original mocap example	Synthesized mocap example
$P(S = 1) = 0.99$	$P(S = 2) = 0.99$

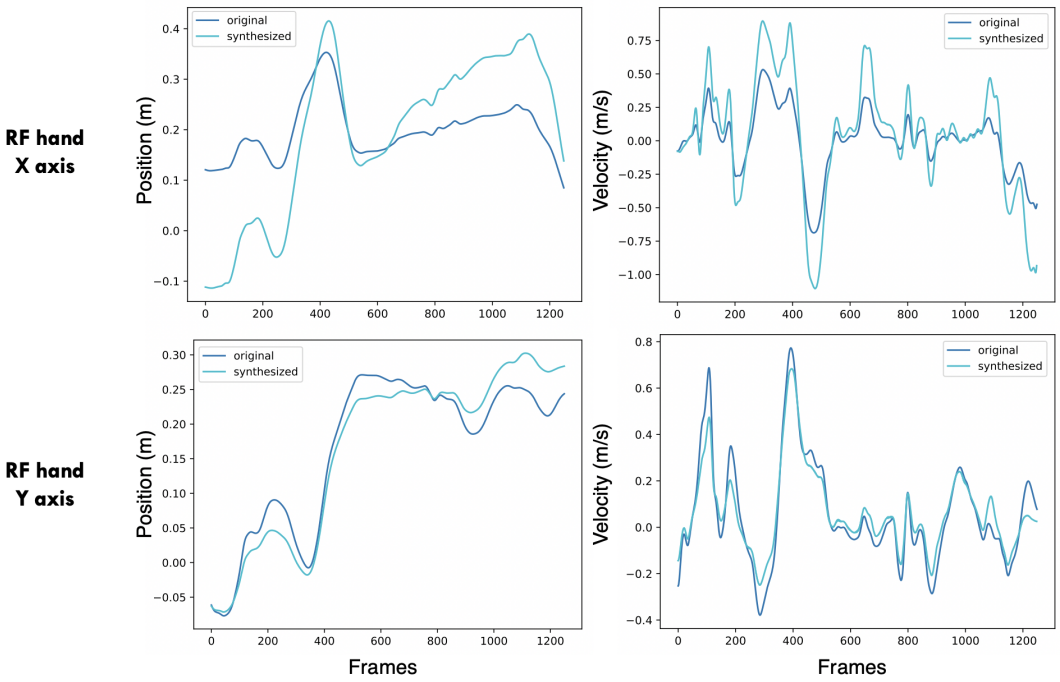


FIGURE 10.9: Position and velocity data of the RF hand marker along X and Y axes, for the original and synthesized movements.

10.2.3 Example 2: anonymization of Signer 1

In this second example, a synthesized mocap excerpt was generated in order to anonymize the movements of Signer 1. For that aim, the synthesis procedure was driven by the following target statistics:

$$\tilde{\boldsymbol{d}}_{-\mathbf{50}} = \boldsymbol{d}_{orig} - 50\boldsymbol{d}_{\mathbf{1}} \tag{10.8}$$

where $\boldsymbol{d}_{orig}$ are the statistics of the mocap example 1 of Signer 1, and $\boldsymbol{d}_{\mathbf{1}}$ are the discriminant statistical patterns of Signer 1.

The synthesis procedure involved 5,000 iterations and a step size of 0.0001. Important kinematic statistics characteristic of Signer 1 were modified through the process (see Figure 10.8). For instance, the SD of position and SD of velocity of both hands were increased along all of the three axes. Moreover, a significant (negative) covariance between the right and left hand across the X axis was generated. A significant (positive) covariance between the two hands also appeared along the Z axis.

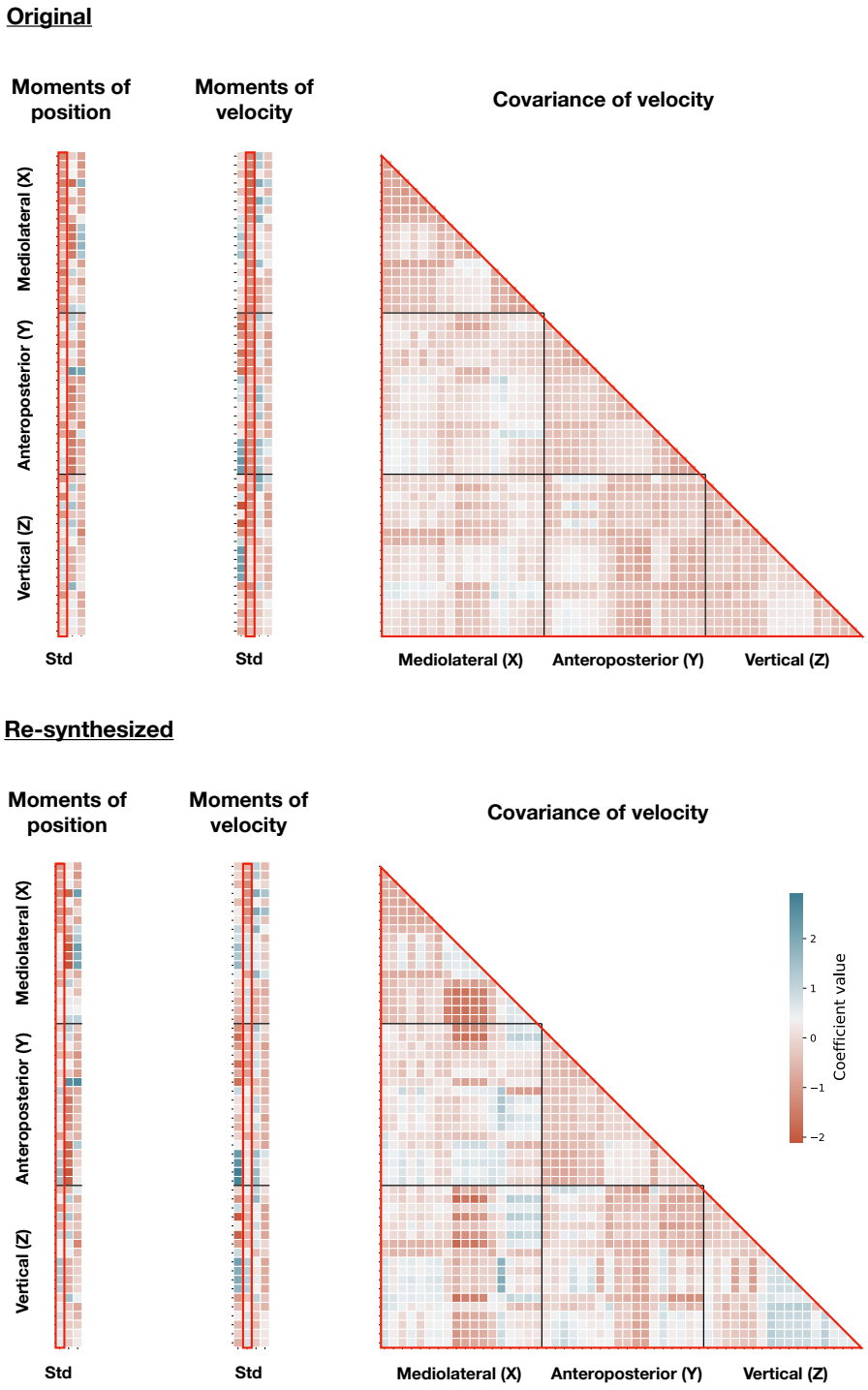


FIGURE 10.10: Statistics of kinematic features of mocap example 1 of Signer 1, before (original) and after (re-synthesized) the synthesis procedure: moments (columns: std, skew, kurtosis) of position for all markers (rows), moments (columns: mean, std, skew, kurtosis) of velocity for all markers (rows), covariance of velocity between markers (rows and columns). Markers are sorted from 1 to 19 as presented in Section 5.3.1 along X, Y and Z axes. Blue represents positive statistical values, while red represents negative ones. The three main classes of statistics that carry identity information are highlighted.

As shown in Figure 10.11 (and Videos 10.5 and 10.6), the overall velocity of the RF hand markers was modified so that the SD of velocity increased, along the three X, Y and Z axes. This is in line with the previous observations made from Figure 10.10. Among other patterns, modifying these aspects of motion may allow Signer 1 to be non-identifiable. This latter hypothesis was confirmed by the automatic signer identification model, which did not manage to identify Signer 1 from the synthesized movements, as shown in Table 10.3:

TABLE 10.3: Output of the automatic signer identification model. P is the probability that the movements were produced by the signer (see Equation 9.4).

Original mocap example	Synthesized mocap example
$P(S = 1) = 0.99$	$P(S = 1) = 0.05$*

* The highest probability was that of Signer 4 ($P(S = 4) = 0.43$).

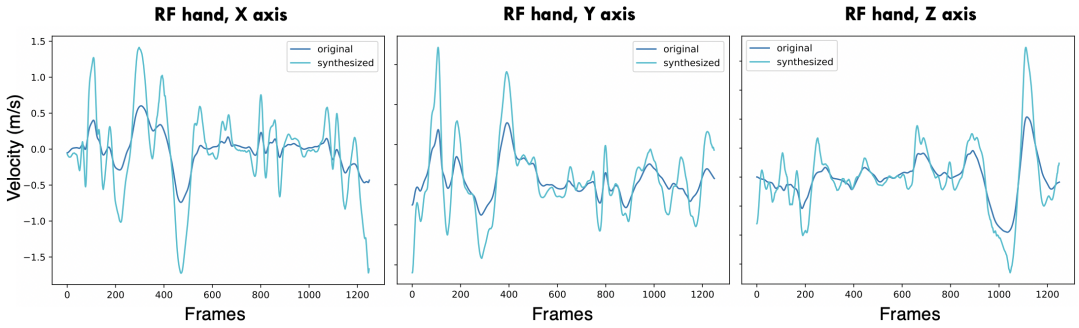


FIGURE 10.11: Velocity data of the RF hand marker along X, Y and Z axes, for the original and synthesized movements.

10.3 Discussion

In Chapter 9, we found that simple time-averaged statistics of kinematic features of the movements could allow for the automatic identification of signers with a high accuracy. In the present chapter, we demonstrated that the statistical patterns characteristic of the identiy of the signers could be manipulated in the movements in order to re-generate original mocap recordings with a modified identity attribute. For that aim, the mocap data of a given SL discourse were modified so that a set of their statistics (i.e., mean of position, mean of velocity, SD of position, SD of velocity and covariance of velocity between markers) reached target values. These target values could be set for various manipulations, such as kinematic identity conversion (from one signer to another) or anonymization. Using a sufficient number of iterations and optimizing values for the step size and the loss weights, the algorithm converged and the statistics of the synthesized movements approached the target values.

When tested on the novel, synthesized, mocap examples, the output of the automatic signer identification model (see Chapter 9) was altered. For instance, exaggerating the kinematic aspects characteristic of Signer 2 in the mocap recording of Signer 1 made this latter identified as movements of Signer 2 (see Section 10.2.2). Moreover, reducing the kinematic aspects characteristic of Signer 1 in one of her

mocap recording made Signer 1 non identifiable (see Section 10.2.3). These computational findings call for further visual perception studies investigating the extent to which the ability of human observers to identify signers would differ between original and synthesized SL movements shown as PLDs. The second outcome of this chapter is that the synthesis algorithm allows modifying the identity attribute in the movements of signers while preserving the original temporal structure of the movements. This is of particular interest for SL research as degrading temporal structure could impair the comprehension of the SL discourse. Again, human experiments would be needed to assess the comprehensibility of SL messages generated from the synthesized movements.

The procedure modifying the statistics of the original movements was driven by the discriminant statistical patterns established for each signer in Chapter 9. The further computational development of this synthesis algorithm provides a tool for visualizing these identity-specific aspects of motion by generating videos, such as PLD ones (see Sections 10.2.2 and 10.2.3). Furthermore, it opens up promising perspectives toward controlling the identity attribute of SL movements in the animation of virtual signers. For now, the synthesis process is achieved with parameters that are optimized manually for each mocap example but further work could ease the optimization process across examples and signers and allow automatically finding the optimal parameters in order to anonymize the SL content of virtual signers.

The general idea of synthesizing new motion while manipulating the identity attribute is to reduce or amplify the importance of discriminant features in the synthesized movements. As previously discussed in Section 4.3, the main challenge of such synthesis tools is to be able to project the identity-specific features back onto the initial high-dimensional motion space. For that aim, studies have used various methods, such as principal movement decomposition (Young and Reinkensmeyer, 2014; Troje, 2002a), Hidden Markov Models (Tilmanne et al., 2012; Tilmanne et al., 2014), gaussian modeling (Tilmanne and Dutoit, 2010) or deep neural networks (Tits, 2018). In the present study, the identity-specific aspects of the movements were carried by statistics. Therefore, we developed a novel synthesis algorithm that allows modifying original movement recordings in order to reach some desired statistics. Similar methods have been previously applied to audio signals (McDermott et al., 2009; McDermott and Simoncelli, 2011; Norman-Haignere and McDermott, 2018) in order to create various sounds from noise sharing the same statistics but with a different fine structure. To the author's knowledge, this study is the first to propose a synthesis method for re-generating mocap recordings while controlling specific statistics of the motion. As already argued in Chapter 9, using statistics as a descriptor for the identity of moving individuals is of particular relevance, as identity is a time-invariant property that humans manage to infer from various discourses, regardless of the semantic content. Moreover, our statistical-based method outperforms temporal, frame-by-frame, methods when applied to mocap recordings that are not synchronized in time across examples and individuals (see Section 9.2.4). Real-life movements are rarely synchronized in time across categories of actions and individuals, even less across discourses in SL. Therefore, further developing statistical-based methods for controlling specific aspects of motion could provide novel tools of interest in addition to the existing temporal-based methods, in particular for controlling human attributes such as identity in virtual signers animations.

Part III Summary

Can signers actually be identified when their movements are replayed via a virtual signer? If so, which features of their movements make them identifiable? Finally, how could these features be manipulated when animating virtual signers in order to control the identity attribute (e.g., to anonymize SL messages)? In this part of the thesis, a visual perception study first demonstrated that signers shown as moving Point-Light Displays (PLDs) could be identified by deaf observers with a significant accuracy, similarly to prior work on non-SL motion. Moreover, the perceptual judgments of the participants were not correlated with cues related to the size and shape of the body of the signers. In order to understand which further aspects of the motion may allow for the identification, we trained a machine learning model to identify signers from statistics of the motion capture (mocap) data. The performance of the model revealed that the identity of signers could be characterized by simple statistics of kinematic aspects of their movements. We finally proposed a synthesis algorithm in order to re-generate mocap recordings while controlling the identity-specific features of the signer. Modified motion examples produced by the algorithm allowed misleading the automatic signer identification model, while maintaining the general structure of the motion. These latter observations calls for further human experiments assessing the extent to which the ability of participants to identify the signers from the synthesized excerpts could be compromised while the comprehensibility of the SL message could remain unaffected. Should this trade-off be reached, our algorithm could allow virtual signers to produce anonymized SL messages, which would open new horizons for deaf SL users.

Conclusions

When technological developments meet fundamental research

As developed in **Chapter 1**, deaf Sign Language (SL) users face many communication barriers. In particular, the vast majority of automatic communication tools are not compatible with SL content, but only with spoken or written one. Developing successful tools for automatic SL processing would allow breaking down these barriers. For that aim, further insights must be gained into multiple disciplines, in particular motion science. Indeed, beyond the sparsity of research and developments conducted in SL compared to spoken languages, the automatic processing of SL is challenging because of the intrinsic complexity of SL movements. For instance, SL involves multiple motion features from various body parts, such as movements of the torso, arms, hands and fingers as well as facial expressions. Moreover, the spatial and temporal coordination of SL gestures is driven by a highly structured linguistic system, whose modeling does not yet meet with a broad consensus among linguists.

One important area of automatic SL processing is SL generation, which aims to automatically produce SL messages using virtual signers, similarly to voice assistants in spoken languages. Promising progress has been made along this line in the past decade, notably with the ability to record the movements of a signer with high accuracy in order to use it for the animation of the virtual signers. Still, many technological developments are needed to provide tools for the automatic generation of SL messages in the same way as for spoken languages. In particular, the present thesis aimed to gain insights into the possibility of anonymizing the movements of a signer, in the same way as a speaker can remain anonymous by modifying specific aspects of the voice. Up to now, no tool allows deaf SL users to remain anonymous when producing a message in SL. Developing computational models able to generate SL animations from the movements of signers while keeping them non-identifiable would allow deaf SL users to further share testimonies that require anonymity or to post comments on forums and social networks directly in SL in the same was as written comments allow for some anonymity in other languages.

For all these reasons, this thesis investigated how computational models could extract human attributes from the movements of individuals and how these attributes could be controlled when generating motion, in the particular case of identity in SL. Beyond the aim of providing novel technological developments that could improve communication tools for deaf SL users, this problem raised fundamental research questions (Part I). First, we questioned the ability of human observers to identify signers from SL movements, as previously shown for non-SL movements such as walking or dancing (**Chapter 2**). We then aimed to determine the critical aspects of the movements that may allow the observers to identify the person, using state-of-the-art 3D motion capture (mocap) systems and computational methods, including machine learning, for motion analysis (**Chapters 3 and 4**).

Main contributions of the thesis

First, the present thesis provided insights into how the complex structure of SL motion in spontaneous discourse could be modeled (Part II), which was crucial for further investigating the encoding of identity information in the movements. Many prior studies have investigated SL motion in isolation (e.g., with isolated signs or fingerspelling) and some of them have focused on only a few aspects of SL motion (e.g., finger gestures), which has shed light on many properties of SL motion but only for a limited subset of the movements produced in real conditions. One key objective of this thesis was to study SL movements within a more realistic framework.

For that aim, we used the **3D mocap recordings of MOCAP1 corpus**, which provides **continuous French Sign Language (LSF)** productions of multiple signers. The mocap data we used in the present thesis consisted of the 3D trajectories of the upper-body markers of six signers who had described 24 pictures in LSF in a spontaneous manner. We then developed **novel motion representations and preprocessing methods**, in particular for visualization purposes and in order to normalize the movements with respect to size, shape and posture of the body of the signers (**Chapter 5**).

The limitations of analyzing SL motion in isolation were further outlined by a **spectral analysis of the mocap data** of spontaneous LSF, which revealed that the kinematic bandwidth of SL may be wider than priorly demonstrated with isolated signs (**Chapter 6**). Combining Power Spectral Density estimation and residual analysis, results showed that a **reasonable bandwidth for our SL mocap data was 0–12 Hz**. The outcome of this reevaluation is twofold. First, it outlined important information in the movements at significantly higher frequencies than prior estimations made on isolated signs (i.e., 0–6 Hz). This suggests that SL movements may involve higher frequencies in real-life conditions. Moreover, the estimated bandwidth presented in this thesis could be used as a reference when modeling SL movements in real-life conditions for application purposes. For instance, mocap data sampled at high frame rates (e.g., 120 or 250 fps) could be low-pass filtered using a 12-Hz cutoff frequency. Interestingly, this 12-Hz value is compatible with the use of motion data extracted from a video, which still is the most frequent type of data used in automatic SL processing, in particular SL recognition. Indeed, the standard frame rates of videos (i.e., 24 fps or higher) allow filtering the data with a 12-Hz frequency (or lower), according to the Nyquist-Shannon theorem (Nyquist, 1928; Shannon, 1949).

Computational models for the automatic processing of SL thus could use a reduced representation of the movements in terms of frequencies. Mocap data of SL, however, remain in a high-dimensional space, because of the high number of body markers they involve. In **Chapter 7**, we tested a **dimensionality reduction technique on the mocap data** in order to assess the extent to which complex upper-body movements of spontaneous SL could be decomposed into elementary movements. Principal Component Analysis of the data revealed that the **SL movements could be reduced to a set of eight Principal Movements (PMs)**, which accounted for 95% of the variance in the motion, both for the six individual signers and across all signers. Unlike in most prior studies investigating PMs of human movements, our mocap data was not synchronized in time across examples and signers. This suggests that despite their complexity, the high-dimensional movements of SL in real-life conditions may be characterized by key movements in a space of lower dimension. For application purposes, these findings could ease the incorporation of dense 3D mocap datasets in models of automatic SL processing, which could use a reduced subset of elementary movements while keeping the information intact.

As developed in Chapter 4, differences in the execution of these PMs between individuals have been successfully used to automatically extract various relevant attributes, such as neuromuscular disorders, gesture expertise or gender. The identity of moving individuals has also been predicted using further descriptions of the movements, such as key postures for gait or kinematic statistics for dancing. Therefore, we further investigated how signers could actually be identified from their motion in SL and how identity information could be encoded in the movements (Part III).

First, a **visual perception study** was conducted where deaf observers were asked to **identify four signers from their movements** in spontaneous LSF, shown as Point-Light Displays (PLDs) (**Chapter 8**). The performance of the participants, jointly with computational analyses of the mocap data, showed that **the movements contained enough information to allow for accurate identifications of the signers**, beyond cues related to the size and shape of the body of the signers. These results suggest that identity information may be carried by further motion features, in particular kinematic ones.

A **machine learning model** was thus trained on statistics of the mocap data for **automatic signer identification**, in order to determine the further aspects of motion that allow inferring the identity of a signer (**Chapter 9**). The performance of the model using normalized mocap data confirmed the minor role of size and shape differences in the identification. A significant effect of the average posture of the signers was found. However, the identification accuracy of the model remained substantial (86.8%) even when having normalized for size, shape and posture of each individual. The kinematic signature of the identity of the signers was further defined by distinct, uncorrelated, statistical patterns used by the model in the identification. Our statistical-based approach outperformed the widely used temporal-based method mentioned above: PM decomposition. The successful extraction of identity information from kinematic statistics confirms the fundamental statement that **human attributes may be carried mostly by the movements of the individuals per se, rather than cues related to the structure of their body in motion**. Moreover, it opens up promising perspectives toward controlling the identity-specific aspects of SL movements in real-life conditions (i.e., where movements are not synchronized in time across individuals and examples, making temporal-based analyses complex) for generation purposes.

To illustrate these potentials for automatic SL generation, we finally proposed a **synthesis algorithm** which successfully allowed **modifying the identity attribute in existing SL mocap recordings** (**Chapter 10**). In the synthesis procedure, the statistics of the mocap example are measured while target statistics are defined in order to reduce or exaggerate the statistical kinematic patterns characteristic of one signer. This method can be used for multiple manipulations, such as anonymization (i.e., reducing the importance of statistics characteristic of the signer) or identity conversion (i.e., exaggerating the importance of statistics characteristic of another signer). The target statistics were approached using an imposing algorithm, which iteratively modified the mocap signals using gradient descent. Moreover, additional constraints in the imposing algorithm allowed us to modify the identity attribute in the movements, while preserving the initial temporal structure of the original movements. These results call for further research investigating how the signers could be misidentified by human observers from the modified motion examples shown as PLDs, and how the comprehensibility of the SL content could remain unaltered.

Future work and perspectives

The present thesis provided original contributions to the fields of motion, computer science and visual perception. In particular, it shed light on how complex SL movements could be modeled in realistic, spontaneous, discourses, and opened up promising perspectives for automatically controlling identity information in the movements for generation applications. More generally, the present machine learning developments for the kinematic control of identity could be of interest in a wider area and for further motion aspects. For instance, similar methods could be applied to emotion recognition in dancing, analysis of aging effects on gait or expertise evaluation in musical gestures. Still, further work is needed and some limitations of the present work must be overcome in order to effectively use these tools in actual applications, such as for anonymizing virtual signers in SL.

First, although we aimed to use SL mocap data as representative as possible of real-life conditions (i.e., spontaneous LSF discourses), the data of MOCAP1-v2 was limited to the movements of six signers. Moreover, the LSF discourses used in the present thesis were picture descriptions, which may have involved specific linguistic structures more than others (e.g., depicting ones). The different outcomes reported from Chapter 6 to Chapter 10 should be further tested with other signers and in a wider linguistic context. It should be noted that most prior SL studies investigated the movements of a lower number of signers, notably because of the difficulty to create accurate 3D mocap corpora with multiple individuals (see Chapter 3). Still, one could expect the performance of automatic identification models to decline as the number of individuals to identify increases, which motivates the need for further tests of our methods on more signers.

Moreover, our studies neither focused on facial expressions nor on finger movements of the signers. Our results involved various upper-body parts (i.e., stomach, sternum, shoulders, elbows, wrists, hands and head), which may carry an important part of the identity information in the movements. Still, signers could produce identity-specific motion features with facial movements and finger gestures. For instance, the performance of our automatic identification model could be assessed using facial mocap recordings including eyebrows, whose motion has been priorly shown to have a role in the intonation of SL discourses. This could be tested using the facial mocap recordings of MOCAP1 (see Chapter 5).

Then, further work should be carried out with the aim of using the presented methods in real-life applications for automatic SL generation. First, some improvement prospects of the synthesis algorithm should be outlined. For instance, the quality of matching of the statistics could be optimized. For now, statistics of the synthesized movements significantly approach the target ones but sometimes with non-negligible approximations, in particular for high-dimensional statistics (e.g., covariance of velocity between markers). Moreover, this first version of the algorithm requires finding the optimal values for the parameters of the imposing procedure manually (i.e., loss weights, step size, number of iterations), which is not well suited for real-time and real-life applications. In order to overcome these limitations, further machine learning techniques could be tested. Furthermore, the extent to which the imposing algorithm manage to approach the target statistics may be limited by the high number of statistics involved. To address that problem, we could build a model that affect higher weights to the statistics of importance than to other ones, in order to reduce the complexity of the imposing procedure. Additionally, further methods could be tested in order to preserve the initial structure of the movements,

as the one used in this thesis (i.e., imposing the correlation between original and synthesized velocities) was quite empirical.

Finally, the main developments of this thesis open up promising application perspectives, which calls for further work investigating how our methods could be applied in real-life conditions. First, the mocap setup involved in the application could be more portable and more accessible than state-of-the-art 3D mocap systems. Compared to the optical *Optitrack* system used in the present thesis, the body trajectories could be extracted from videos of markerless RGB or depth cameras, using image processing techniques to recover the missing 3D information (Cao et al., 2019; Belissen et al., 2020). This would allow testing our methods on a wider variety of motion recordings and signers and could be useful to make this automatic SL processing application accessible to the general public (e.g., for smartphone applications). Furthermore, the synthesis algorithm for kinematic control of identity needs to be tested with human participants. We should investigate three key problems: (1) identifiability, by verifying that the ability of human observers to identify the signers is compromised when showing the synthesized modified movements, as compared to the original ones; (2) comprehensibility, by evaluating the extent to which the observers still understand the SL content in the modified motion examples; and (3) acceptability, by assessing the deaf user perspective on the virtual signers animated with the modified movements and discussing potential use cases (e.g., with focus groups). Should these three fundamental points be validated, the present thesis could constitute a first step of interest toward automatically controlling the identity of deaf SL users when expressing themselves via virtual signers. In particular, this could allow developing effective applications for the production of fully-anonymized SL messages.

Bibliography

"Jade". URL: https://www.accessibilite.sncf.com/la-lettre-de-l-accessibilite/lettres/2010/decembre-2010-no2/article/jade-des-mots-pleins-les-mains.

"Keia". URL: https://www.keia.io/.

Abdi, Hervé and Lynne J Williams (2010). "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4, pp. 433–459.

Agus, Trevor R, Clara Suied, Simon J Thorpe, and Daniel Pressnitzer (2012). "Fast recognition of musical sounds based on timbre". In: *The Journal of the Acoustical Society of America* 131.5, pp. 4124–4133.

Alemi, Omid, William Li, and Philippe Pasquier (2015). "Affect-expressive movement generation with factored conditional restricted boltzmann machines". In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 442–448.

Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. (2016). "Deep speech 2: End-to-end speech recognition in english and mandarin". In: *International conference on machine learning*. PMLR, pp. 173–182.

Atkinson, Anthony P, Winand H Dittrich, Andrew J Gemmell, and Andrew W Young (2004). "Emotion perception from dynamic and static body expressions in point-light and full-light displays". In: *Perception* 33.6, pp. 717–746.

Babaee, Maryam, Linwei Li, and Gerhard Rigoll (2019). "Person identification from partial gait cycle using fully convolutional neural networks". In: *Neurocomputing* 338, pp. 116–125.

Baragchizadeh, Asal, Parisa R Jesudasen, and Alice J O'Toole (2020). "Identification of unfamiliar people from point-light biological motion: A perceptual reevaluation". In: *Visual Cognition* 28.9, pp. 513–522.

Beardsworth, T and T Buckner (1981). "The ability to recognize oneself from a video recording of one's movements without seeing one's body". In: *Bulletin of the Psychonomic Society* 18.1, pp. 19–22.

Belissen, Valentin, Annelies Braffort, and Michèle Gouiffès (2020). "Dicta-Sign-LSF-v2: remake of a continuous French sign language dialogue corpus and a first baseline for automatic sign language processing". In: *LREC 2020, 12th Conference on Language Resources and Evaluation*.

Benchiheub, Mohamed-El-Fataf, Annelies Braffort, Bastien Berret, and Cyril Verrecchia (2016a). *MOCAP1*. https://www.ortolang.fr/market/item/mocap1. Limsi, distributed via ORTOLANG (Open Re-sources and TOols for LANGuage).

Benchiheub, Mohamed-El-Fatah (2017). "Contribution à l'analyse des mouvements 3D de la Langue des Signes Française (LSF) en Action et en Perception". PhD thesis. Université Paris-Saclay (ComUE).

Benchiheub, Mohamed-el-Fatah, Bastien Berret, and Annelies Braffort (Jan. 2016b). "Collecting and Analysing a Motion-Capture Corpus of French Sign Language".

In: *Workshop on the Representation and Processing of Sign Languages*. Portoroz, Slovenia. URL: https://hal.archives-ouvertes.fr/hal-01633625.

Bernstein, Nik (1927). "Kymozyklographion, ein neuer Apparat für Bewegungsstudium". In: *Pflüger's Archiv für die gesamte Physiologie des Menschen und der Tiere* 217.1, pp. 782–792.

Bernstein, Nikolai (1966). "The co-ordination and regulation of movements". In: *The co-ordination and regulation of movements*.

Berret, Bastien, François Bonnetblanc, Charalambos Papaxanthis, and Thierry Pozzo (2009). "Modular control of pointing beyond arm's length". In: *Journal of Neuroscience* 29.1, pp. 191–205.

Berret, Bastien, Christian Darlot, Frédéric Jean, Thierry Pozzo, Charalambos Papaxanthis, and Jean Paul Gauthier (2008). "The inactivation principle: mathematical solutions minimizing the absolute work and biological implications for the planning of arm movements". In: *PLoS computational biology* 4.10, e1000194.

Berret, Bastien and Frédéric Jean (2016). "Why don't we move slower? the value of time in the neural control of action". In: *Journal of neuroscience* 36.4, pp. 1056–1070.

Bertenthal, Bennett I and Jeannine Pinto (1994). "Global processing of biological motions". In: *Psychological science* 5.4, pp. 221–225.

Bertenthal, Bennett I, Dennis R Proffitt, and James E Cutting (1984). "Infant sensitivity to figural coherence in biomechanical motions". In: *Journal of experimental child psychology* 37.2, pp. 213–230.

Bertenthal, Bennett I, Dennis R Proffitt, and Steven J Kramer (1987). "Perception of biomechanical motions by infants: implementation of various processing constraints." In: *Journal of Experimental Psychology: Human Perception and Performance* 13.4, p. 577.

Bevilacqua, Frederic, Fabrice Guédy, Norbert Schnell, Emmanuel Fléty, and Nicolas Leroy (2007). "Wireless sensor interface and gesture-follower for music pedagogy". In: *Proceedings of the 7th international conference on New interfaces for musical expression*, pp. 124–129.

Bigand, Emmanuel, Charles Delbé, Yannick Gérard, and Barbara Tillmann (2011). "Categorization of extremely brief auditory stimuli: domain-specific or domain-general processes?" In: *PloS one* 6.10, e27024.

Bigand, Félix, Elise Prigent, Bastien Berret, and Annelies Braffort (2021a). "Decomposing spontaneous sign language into elementary movements: A principal component analysis-based approach". In: *PloS one* 16.10, e0259464.

— (2021b). "How fast is Sign Language? A reevaluation of the kinematic bandwidth using motion capture". In: *To be published in: Proceedings of the 29th European Signal Processing Conference*.

— (2021c). "Machine learning of motion statistics reveals the kinematic signature of a person's identity in sign language". In: *Frontiers in Bioengineering and Biotechnology* 9, p. 603.

Bigand, Félix, Elise Prigent, and Annelies Braffort (2020). "Person Identification Based on Sign Language Motion: Insights from Human Perception and Computational Modeling". In: *Proceedings of the 7th International Conference on Movement and Computing*, pp. 1–7.

Bishop, Gary, Greg Welch, and B Danette Allen (2001). "Tracking: Beyond 15 minutes of thought". In: *SIGGRAPH Course Pack* 11.

Blakemore, Sarah-Jayne and Jean Decety (2001). "From the perception of action to the understanding of intention". In: *Nature reviews neuroscience* 2.8, pp. 561–567.

Blanz, Volker and Thomas Vetter (1999). "A morphable model for the synthesis of 3D faces". In: *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 187–194.

Bläsing, Bettina E and Odile Sauzet (2018). "My action, my self: Recognition of self-created but visually unfamiliar dance-like actions from point-light displays". In: *Frontiers in psychology* 9, p. 1909.

Blondel, Marion, Dominique Boutet, Fanny Catteau, and Coralie Vincent (2019). "Signing amplitude and other prosodic cues in older signers: Insights from motion capture from the SignAge Corpus". In: *Corpora for Language and Aging Research (CLARe 4)*.

Bobick, Aaron F. and James W. Davis (2001). "The recognition of human movement using temporal templates". In: *IEEE Transactions on pattern analysis and machine intelligence* 23.3, pp. 257–267.

Braffort, Annelies (1996). "A gesture recognition architecture for sign language". In: *Proceedings of the second annual ACM conference on Assistive technologies*, pp. 102–109.

Braffort, Annelies, Laurence Bolot, and Jérémie Segouat (2011). "Virtual signer coarticulation in Octopus, a Sign Language generation platform". In: *GW 2011: The 9th International Gesture Workshop*.

Brand, Matthew and Aaron Hertzmann (2000). "Style machines". In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 183–192.

Brashear, Helene, Thad Starner, Paul Lukowicz, and Holger Junker (2003). "Using multiple sensors for mobile sign language recognition". In: Georgia Institute of Technology.

Brownlow, Sheila, Amy R Dixon, Carrie A Egbert, and Rebecca D Radcliffe (1997). "Perception of movement and dancer characteristics from point-light displays of dance". In: *The Psychological Record* 47.3, pp. 411–422.

Bruce, Vicki, Tim Valentine, and Alan Baddeley (1987). "The basis of the 3/4 view advantage in face recognition". In: *Applied cognitive psychology* 1.2, pp. 109–120.

Bülthoff, Isabelle, Heinrich Bülthoff, and Pawan Sinha (1998). "Top-down influences on stereoscopic depth-perception". In: *Nature neuroscience* 1.3, pp. 254–257.

Camgoz, Necati Cihan, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden (2018). "Neural sign language translation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7784–7793.

Camurri, Antonio, Ingrid Lagerlöf, and Gualtiero Volpe (2003a). "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques". In: *International journal of human-computer studies* 59.1-2, pp. 213–225.

Camurri, Antonio, Barbara Mazzarino, and Gualtiero Volpe (2003b). "Analysis of expressive gesture: The eyesweb expressive gesture processing library". In: *International gesture workshop*. Springer, pp. 460–467.

Cao, Zhe, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh (2019). "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields". In: *IEEE transactions on pattern analysis and machine intelligence* 43.1, pp. 172–186.

Carlson, Emily, Pasi Saari, Birgitta Burger, and Petri Toiviainen (2020). "Dance to your own drum: Identification of musical genre and individual dancer from motion capture using machine learning". In: *Journal of New Music Research*, pp. 1–16.

Catteau, Fanny, Marion Blondel, Coralie Vincent, Patrice Guyot, and Dominique Boutet (2016). "Variation prosodique et traduction poétique (LSF/français): Que

devient la prosodie lorsqu'elle change de canal?(Prosodic variation and poetic translation (LSF/French): What happens to prosody with a channel change?)[In French]". In: *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 1: JEP*, pp. 750–758.

Chai, Xiujuan, Hanjie Wang, and Xilin Chen (2014). "The devisign large vocabulary of chinese sign language database and baseline evaluations". In: *Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS*.

Charayaphan, C and AE Marble (1992). "Image processing system for interpreting motion in American Sign Language". In: *Journal of Biomedical Engineering* 14.5, pp. 419–425.

Cherniavsky, Neva, Anna C Cavender, Richard E Ladner, and Eve A Riskin (2007). "Variable frame rate for low power mobile sign language communication". In: *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pp. 163–170.

Cooper, HM, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden (2012). "Sign language recognition using sub-units". In: *Journal of Machine Learning Research* 13, pp. 2205–2231.

Corballis, Michael C (1999). "The Gestural Origins of Language: Human language may have evolved from manual gestures, which survive today as a" behavioral fossil" coupled to speech". In: *American Scientist* 87.2, pp. 138–145.

Corina, David P, Susan L McBurney, Carl Dodrill, Kevin Hinshaw, Jim Brinkley, and George Ojemann (1999). "Functional roles of Broca's area and SMG: evidence from cortical stimulation mapping in a deaf signer". In: *Neuroimage* 10.5, pp. 570–581.

Crasborn, Onno A and IEP Zwitserlood (2008). "The Corpus NGT: an online corpus for professionals and laymen". In:

Cristianini, Nello, John Shawe-Taylor, et al. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Cui, Runpeng, Hu Liu, and Changshui Zhang (2019). "A deep neural framework for continuous sign language recognition by iterative training". In: *IEEE Transactions on Multimedia* 21.7, pp. 1880–1891.

Cunado, David, Mark S Nixon, and John N Carter (1997). "Using gait as a biometric, via phase-weighted magnitude spectra". In: *International conference on audio-and video-based biometric person authentication*. Springer, pp. 93–102.

— (2003). "Automatic extraction and description of human gait models for recognition purposes". In: *Computer vision and image understanding* 90.1, pp. 1–41.

Cutting, James E and Lynn T Kozlowski (1977). "Recognizing friends by their walk: Gait perception without familiarity cues". In: *Bulletin of the psychonomic society* 9.5, pp. 353–356.

Cutting, James E, Cassandra Moore, and Roger Morrison (1988). "Masking the motions of human gait". In: *Perception & psychophysics* 44.4, pp. 339–347.

Cutting, James E, Dennis R Proffitt, and Lynn T Kozlowski (1978). "A biomechanical invariant for gait perception." In: *Journal of Experimental Psychology: Human Perception and Performance* 4.3, p. 357.

Dahl, Sofia and Anders Friberg (2007). "Visual perception of expressiveness in musicians' body movements". In: *Music Perception* 24.5, pp. 433–454.

Dalal, Navneet and Bill Triggs (2005). "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee, pp. 886–893.

Davidson, Jane W (1993). "Visual perception of performance manner in the movements of solo musicians". In: *Psychology of music* 21.2, pp. 103–113.

Davydov, Maksym and Olga Lozynska (2017). "Information system for translation into Ukrainian sign language on mobile devices". In: *2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*. Vol. 1. IEEE, pp. 48–51.

De Cheveigné, Alain and Jonathan Z Simon (2007). "Denoising based on time-shift PCA". In: *Journal of neuroscience methods* 165.2, pp. 297–305.

— (2008). "Denoising based on spatial filtering". In: *Journal of neuroscience methods* 171.2, pp. 331–339.

de'Sperati, Claudio and Paolo Viviani (1997). "The relationship between curvature and velocity in two-dimensional smooth pursuit eye movements". In: *Journal of Neuroscience* 17.10, pp. 3932–3945.

Dilsizian, Mark, Zhiqiang Tang, Dimitris Metaxas, Matt Huenerfauth, and Carol Neidle (2016). "The importance of 3D motion trajectories for computer-based sign recognition". In: *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining. LREC 2016, Portorož, Slovenia. May 2016*.

Dittrich, Winand H (1993). "Action categories and the perception of biological motion". In: *Perception* 22.1, pp. 15–22.

Dittrich, Winand H, Tom Troscianko, Stephen EG Lea, and Dawn Morgan (1996). "Perception of emotion from dynamic point-light displays represented in dance". In: *Perception* 25.6, pp. 727–738.

Duarte, Kyle and Sylvie Gibet (2010). "Heterogeneous data sources for signed language analysis and synthesis: The signcom project". In: *Seventh international conference on Language Resources and Evaluation (LREC 2010)*. Vol. 2. European Language Resources Association, pp. 1–8.

Ebling, Sarah and John Glauert (2016). "Building a Swiss German Sign Language avatar with JASigning and evaluating it among the Deaf community". In: *Universal Access in the Information Society* 15.4, pp. 577–587.

Efthimiou, Eleni, Stavroula-Evita Fontinea, Thomas Hanke, John Glauert, Rihard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Goudenove (2010). "Dicta-sign–sign language recognition, generation and modelling: a research effort with applications in deaf communication". In: *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pp. 80–83.

Elliott, Ralph, John RW Glauert, JR Kennaway, and Ian Marshall (2000). "The development of language processing support for the ViSiCAST project". In: *Proceedings of the fourth international ACM conference on Assistive technologies*, pp. 101–108.

Emmorey, Karen (2001). *Language, cognition, and the brain: Insights from sign language research*. Psychology Press.

Emmorey, Karen, Robin Thompson, and Rachael Colvin (2009). "Eye gaze during comprehension of American Sign Language by native and beginning signers". In: *Journal of Deaf Studies and Deaf Education* 14.2, pp. 237–243.

Erard, Michael (2017). *"Why Sign-Language Gloves Don't Help Deaf People."* Ed. by The Atlantic. URL: https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/.

Federolf, PA, KA Boyer, and TP Andriacchi (2013a). "Application of principal component analysis in clinical gait research: identification of systematic differences between healthy and medial knee-osteoarthritic gait". In: *Journal of biomechanics* 46.13, pp. 2173–2178.

Federolf, Peter, Robert Reid, Matthias Gilgien, Per Haugen, and Gerald Smith (2014). "The application of principal component analysis to quantify technique in sports". In: *Scandinavian journal of medicine & science in sports* 24.3, pp. 491–499.

Federolf, Peter, Lilian Roos, and Benno M Nigg (2013b). "Analysis of the multi-segmental postural movement strategies utilized in bipedal, tandem and one-leg stance as quantified by a principal component decomposition of marker coordinates". In: *Journal of biomechanics* 46.15, pp. 2626–2633.

Federolf, Peter A (2016). "A novel approach to study human posture control:"Principal movements" obtained from a principal component analysis of kinematic marker data". In: *Journal of biomechanics* 49.3, pp. 364–370.

Felis, Martin L, Katja Mombaur, and Alain Berthoz (2015). "An optimal control approach to reconstruct human gait dynamics from kinematic data". In: *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, pp. 1044–1051.

Fels, Sidney S and Geoffrey E Hinton (1993). "Glove-talk: A neural network interface between a data-glove and a speech synthesizer". In: *IEEE transactions on Neural Networks* 4.1, pp. 2–8.

Filhol, Michael and John Mcdonald (2020). "The Synthesis of Complex Shape Deployments in Sign Language". In: *Workshop on the Representation and Processing of Sign Languages*.

Filhol, Michael, John Mcdonald, and Rosalee Wolfe (2017). "Synthesizing sign language by connecting linguistically structured descriptions to a multi-track animation system". In: *International Conference on Universal Access in Human-Computer Interaction*. Springer, pp. 27–40.

Fitts, Paul M (1954). "The information capacity of the human motor system in controlling the amplitude of movement." In: *Journal of experimental psychology* 47.6, p. 381.

Forster, Jens, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney (2014). "Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather." In: *LREC*, pp. 1911–1916.

Foulds, Richard A (2004). "Biomechanical and perceptual constraints on the bandwidth requirements of sign language". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 12.1, pp. 65–72.

Fox, Robert and Cynthia McDaniel (1982). "The perception of biological motion by human infants". In: *Science* 218.4571, pp. 486–487.

Freund, H-J and Hans-Joachim Büdingen (1978). "The relationship between speed and amplitude of the fastest voluntary contractions of human arm muscles". In: *Experimental Brain Research* 31.1, pp. 1–12.

Furuya, Shinichi, Martha Flanders, and John F Soechting (2011). "Hand kinematics of piano playing". In: *Journal of neurophysiology* 106.6, pp. 2849–2864.

Gachoud, Jean-Pierre, Pierre Mounoud, Claude-Alain Hauert, and Paolo Viviani (1983). "Motor strategies in lifting movements: a comparison of adult and child performance". In: *Journal of Motor Behavior* 15.3, pp. 202–216.

Gibet, Sylvie (2018). "Building french sign language motion capture corpora for signing avatars". In: *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC 2018*.

Grèzes, Julie, Christopher D Frith, and Richard E Passingham (2004). "Inferring false beliefs from the actions of oneself and others: an fMRI study". In: *Neuroimage* 21.2, pp. 744–750.

Grimes, Gary J (Nov. 1983). *Digital data entry glove interface device*. US Patent 4,414,537.

Grobel, Kirsti and Marcell Assan (1997). "Isolated sign language recognition using hidden Markov models". In: *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*. Vol. 1. IEEE, pp. 162–167.

Grossman, ED and R Blake (2001). "Brain activity evoked by inverted and imagined biological motion". In: *Vision research* 41.10-11, pp. 1475–1482.

Grossman, Emily, Michael Donnelly, R Price, D Pickens, V Morgan, G Neighbor, and Randolph Blake (2000). "Brain areas involved in perception of biological motion". In: *Journal of cognitive neuroscience* 12.5, pp. 711–720.

Guitteny, Pierre (2006). "Le passif en langue des signes". PhD thesis. Université Michel de Montaigne-Bordeaux III.

Guo, Guodong, Stan Z Li, and Kapluk Chan (2000). "Face recognition by support vector machines". In: *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. no. PR00580)*. IEEE, pp. 196–201.

Gypsy 7. URL: https://metamotion.com/gypsy/gypsy-motion-capture-system.htm.

Haid, Thomas H, Aude-Clémence M Doix, Benno M Nigg, and Peter A Federolf (2018). "Age effects in postural control analyzed via a principal component analysis of kinematic data and interpreted in relation to predictions of the optimal feedback control theory". In: *Frontiers in aging neuroscience* 10, p. 22.

Hamilton, William Rowan (1866). *Elements of quaternions*. Longmans, Green, & Company.

Hanke, Thomas (2004). "HamNoSys-representing sign language data in language resources and language processing contexts". In: *LREC*. Vol. 4, pp. 1–6.

Hebb, Donald Olding (1949). *The organisation of behaviour: a neuropsychological theory*. Science Editions New York.

Heloir, Alexis, Sylvie Gibet, Franck Multon, and Nicolas Courty (2005). "Captured motion data processing for real time synthesis of sign language". In: *International Gesture Workshop*. Springer, pp. 168–171.

Hickok, Gregory, Mark Kritchevsky, Ursula Bellugi, and Edward S Klima (1996). "The role of the left frontal operculum in sign language aphasia". In: *Neurocase* 2.5, pp. 373–380.

Hietanen, Jari K, Jukka M Leppänen, and Ulla Lehtonen (2004). "Perception of emotions in the hand movement quality of Finnish sign language". In: *Journal of nonverbal behavior* 28.1, pp. 53–64.

Holt, Judith A (1993). "Stanford Achievement Test—8th edition: Reading comprehension subgroup results". In: *American Annals of the Deaf* 138.2, pp. 172–175.

Huang, Jie, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li (2018). "Video-based sign language recognition without temporal segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.

Huang, Ping S, Chris J Harris, and Mark S Nixon (1999). "Recognising humans by gait via parametric canonical space". In: *Artificial Intelligence in Engineering* 13.4, pp. 359–366.

Huynh-The, Thien, Cam-Hao Hua, Nguyen Anh Tu, and Dong-Seong Kim (2020). "Learning 3D spatiotemporal gait feature by convolutional network for person identification". In: *Neurocomputing* 397, pp. 192–202.

Iacoboni, Marco, Roger P Woods, Marcel Brass, Harold Bekkering, John C Mazziotta, and Giacomo Rizzolatti (1999). "Cortical mechanisms of human imitation". In: *science* 286.5449, pp. 2526–2528.

Jacobs, Alissa, Jeannine Pinto, and Maggie Shiffrar (2004). "Experience, context, and the visual perception of human movement." In: *Journal of Experimental Psychology: Human Perception and Performance* 30.5, p. 822.

Jacobs, Alissa and Maggie Shiffrar (2005). "Walking perception by walking observers." In: *Journal of Experimental Psychology: Human Perception and Performance* 31.1, p. 157.

Jantunen, Tommi, Birgitta Burger, Danny De Weerdt, Irja Seilola, and Tuija Wainio (2012). "Experiences from collecting motion capture data on continuous signing". In: *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pp. 75–82.

Johansson, Gunnar (1973). "Visual perception of biological motion and a model for its analysis". In: *Perception & psychophysics* 14.2, pp. 201–211.

— (1976). "Spatio-temporal differentiation and integration in visual motion perception". In: *Psychological research* 38.4, pp. 379–393.

Johnston, Trevor et al. (2009). "Creating a corpus of Auslan within an Australian national corpus". In: *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*.

Joze, Hamid Reza Vaezi and Oscar Koller (2018). "Ms-asl: A large-scale data set and benchmark for understanding american sign language". In: *arXiv preprint arXiv:1812.01053*.

Karpouzis, Kostas, George Caridakis, S-E Fotinea, and Eleni Efthimiou (2007). "Educational resources and implementation of a Greek sign language synthesis architecture". In: *Computers & Education* 49.1, pp. 54–74.

Kay, Bruce A, Michael T Turvey, and Onno G Meijer (2003). "An early oscillator model: studies on the biodynamics of the piano strike (Bernstein & Popova, 1930)". In: *MOTOR CONTROL-CHAMPAIGN-* 7.1, pp. 1–45.

Kipp, Michael (2001). "Anvil-a generic annotation tool for multimodal dialogue". In: *Seventh European Conference on Speech Communication and Technology*.

Kipp, Michael, Alexis Heloir, and Quan Nguyen (2011). "Sign language avatars: Animation and comprehensibility". In: *International Workshop on Intelligent Virtual Agents*. Springer, pp. 113–126.

Klima, Edward S and Ursula Bellugi (1979). *The signs of language*. Harvard University Press.

Knoblich, Günther and Rüdiger Flach (2001). "Predicting the effects of actions: Interactions of perception and action". In: *Psychological science* 12.6, pp. 467–472.

Koech, Chemuttaai C (2006). "A kinematic analysis of sign language". In:

Koller, Oscar (2020). "Quantitative survey of the state of the art in sign language recognition". In: *arXiv preprint arXiv:2008.09918*.

Koller, Oscar, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden (2019). "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos". In: *IEEE transactions on pattern analysis and machine intelligence* 42.9, pp. 2306–2320.

Koller, Oscar, O Zargaran, Hermann Ney, and Richard Bowden (2016). "Deep sign: Hybrid CNN-HMM for continuous sign language recognition". In: *Proceedings of the British Machine Vision Conference 2016*.

Koller, Oscar, Sepehr Zargaran, and Hermann Ney (2017). "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4297–4305.

Koller, Oscar, Sepehr Zargaran, Hermann Ney, and Richard Bowden (2018). "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs". In: *International Journal of Computer Vision* 126.12, pp. 1311–1325.

Kozlowski, Lynn T and James E Cutting (1977). "Recognizing the sex of a walker from a dynamic point-light display". In: *Perception & psychophysics* 21.6, pp. 575–580.

Kuhn, Roland, J-C Junqua, Patrick Nguyen, and Nancy Niedzielski (2000). "Rapid speaker adaptation in eigenvoice space". In: *IEEE Transactions on Speech and Audio Processing* 8.6, pp. 695–707.

Lacquaniti, Francesco, Carlo Terzuolo, and Paolo Viviani (1983). "The law relating the kinematic and figural aspects of drawing movements". In: *Acta psychologica* 54.1-3, pp. 115–130.

Lagerlöf, Ingrid and Marie Djerf (2009). "Children's understanding of emotion in dance". In: *European Journal of Developmental Psychology* 6.4, pp. 409–431.

Lang, Simon, Marco Block, and Raúl Rojas (2012). "Sign language recognition using kinect". In: *International Conference on Artificial Intelligence and Soft Computing*. Springer, pp. 394–402.

Latinus, Marianne and Pascal Belin (2011). "Human voice perception". In: *Current Biology* 21.4, R143–R145.

Lewis, Michael (1999). "Social cognition and the self". In: *Early social cognition: Understanding others in the first months of life*, pp. 81–98.

Liang, Rung-Huei and Ming Ouhyoung (1998). "A real-time continuous gesture recognition system for sign language". In: *Proceedings third IEEE international conference on automatic face and gesture recognition*. IEEE, pp. 558–567.

Little, James and Jeffrey Boyd (1998). "Recognizing people by their gait: the shape of motion". In: *Videre: Journal of computer vision research* 1.2, pp. 1–32.

Liu, Wu, Cheng Zhang, Huadong Ma, and Shuangqun Li (2018). "Learning efficient spatial-temporal gait features with deep learning for human identification". In: *Neuroinformatics* 16.3, pp. 457–471.

Longo, Alessia, Thomas Haid, Ruud Meulenbroek, and Peter Federolf (2019). "Biomechanics in posture space: Properties and relevance of principal accelerations for characterizing movement control". In: *Journal of biomechanics* 82, pp. 397–403.

Loula, Fani, Sapna Prasad, Kent Harber, and Maggie Shiffrar (2005). "Recognizing people from their movement." In: *Journal of Experimental Psychology: Human Perception and Performance* 31.1, p. 210.

Lu, Pengfei and Matt Huenerfauth (2010). "Collecting a motion-capture corpus of American Sign Language for data-driven generation research". In: *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pp. 89–97.

— (2014). "Collecting and evaluating the CUNY ASL corpus for research on American Sign Language animation". In: *Computer Speech & Language* 28.3, pp. 812–831.

Lu, Wei-Lwun and James J Little (2006). "Simultaneous tracking and action recognition using the pca-hog descriptor". In: *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*. IEEE, pp. 6–6.

Luck, Geoff, Suvi Saarikallio, and Petri Toiviainen (2009). "Personality traits correlate with characteristics of music-induced movement". In: *ESCOM 2009: 7th Triennial Conference of European Society for the Cognitive Sciences of Music*.

Luck, Geoff, Petri Toiviainen, and Marc R Thompson (2010). "Perception of expression in conductors' gestures: A continuous response study". In: *Music Perception* 28.1, pp. 47–57.

Malaia, Evguenia, John Borneman, and Ronnie B Wilbur (2008). "Analysis of ASL motion capture data towards identification of verb type". In: *Semantics in text processing. STEP 2008 Conference Proceedings*, pp. 155–164.

Malaia, Evie and Ronnie B Wilbur (2012). "Kinematic signatures of telic and atelic events in ASL predicates". In: *Language and speech* 55.3, pp. 407–421.

Malaia, Evie, Ronnie B Wilbur, and Marina Milković (2013). "Kinematic parameters of signed verbs". In: *Journal of Speech, Language, and Hearing Research* 56.5, pp. 1677–1688.

Marey, Etienne-Jules (1874). *Animal mechanism: a treatise on terrestrial and aerial locomotion*. Vol. 11. Henry S. King & Company.

Martínez, Aleix M, Ronnie B Wilbur, Robin Shay, and Avinash C Kak (2002). "Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language". In: *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE, pp. 167–172.

Massey, Joe T, Joseph T Lurito, Giuseppe Pellizzer, and Apostolos P Georgopoulos (1992). "Three-dimensional drawings in isometric conditions: relation between geometry and kinematics". In: *Experimental Brain Research* 88.3, pp. 685–690.

Mather, George and Linda Murdoch (1994). "Gender discrimination in biological motion displays based on dynamic cues". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 258.1353, pp. 273–279.

McDermott, Josh H, Andrew J Oxenham, and Eero P Simoncelli (2009). "Sound texture synthesis via filter statistics". In: *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 297–300.

McDermott, Josh H, Michael Schemitsch, and Eero P Simoncelli (2013). "Summary statistics in auditory perception". In: *Nature neuroscience* 16.4, pp. 493–498.

McDermott, Josh H and Eero P Simoncelli (2011). "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis". In: *Neuron* 71.5, pp. 926–940.

McDonald, John, Rosalee Wolfe, Ronnie B Wilbur, Robyn Moncrief, Evie Malaia, Sayuri Fujimoto, Souad Baowidan, and Jessika Stec (2016). "A new tool to facilitate prosodic analysis of motion capture data and a data-driven technique for the improvement of avatar motion". In: *Proceedings of Language Resources and Evaluation Conference (LREC)*, pp. 153–159.

Méary, David, Elenitsa Kitromilides, Karine Mazens, Christian Graff, and Edouard Gentaz (2007). "Four-day-old human neonates look longer at non-biological motions of a single point-of-light". In: *PloS one* 2.1, e186.

Meltzoff, Andrew N and M Keith Moore (1995). "A Theory of the Role of Imitation in". In: *The self in infancy: Theory and research*, p. 73.

Menache, Alberto (2000). *Understanding motion capture for computer animation and video games*. Morgan kaufmann.

Meurant, Laurence and Aurélie Sinte (2016). "The French Belgian Sign Language Corpus. A User-Friendly Corpus Searchable Online". In: *Proceedings of the 7th workshop on the Representation and Processing of Sign Languages: Corpus Mining: LREC 2016*, pp. 166–174.

Michel, F (1971). "Etude experimentale de la vitesse du geste graphique". In: *Neuropsychologia* 9.1, pp. 1–13.

Morel, Marion, Richard Kulpa, Anthony Sorel, Catherine Achard, and Séverine Dubuisson (2016). "Automatic and generic evaluation of spatial and temporal errors in sport motions". In: *11th International Conference on Computer Vision Theory and Applications (VISAPP 2016)*, pp. 542–551.

Morford, Jill P and Martina L Carlson (2011). "Sign perception and recognition in non-native signers of ASL". In: *Language learning and development* 7.2, pp. 149–168.

Morford, Jill P, Angus B Grieve-Smith, James MacFarlane, Joshua Staley, and Gabriel Waters (2008). "Effects of language experience on the perception of American Sign Language". In: *Cognition* 109.1, pp. 41–53.

Muir, Laura J and Iain EG Richardson (2005). "Perception of sign language and its application to visual communications for deaf people". In: *Journal of Deaf studies and Deaf education* 10.4, pp. 390–401.

Murase, Hiroshi and Rie Sakai (1996). "Moving object recognition in eigenspace representation: gait analysis and lip reading". In: *Pattern recognition letters* 17.2, pp. 155–162.

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press.

Muybridge, Eadweard (1887). *Animal locomotion*. Da Capo Press.

MVN Animate. URL: https://www.xsens.com/products/mvn-animate.

Naert, Lucie, Caroline Larboulette, and Sylvie Gibet (2020). "LSF-ANIMAL: a motion capture corpus in French sign language designed for the animation of signing avatars". In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6008–6017.

Narayan P. Subramaniyam, Lab Talk - Sapien Labs (2018). *Factors that Impact Power Spectral Density Estimation*. URL: https://sapienlabs.co/factors-that-impact-power-spectrum-density-estimation/.

Neidle, Carol, Ashwin Thangali, and Stan Sclaroff (2012). "Challenges in development of the american sign language lexicon video dataset (asllvd) corpus". In: *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*. Citeseer.

Neri, Peter, M Concetta Morrone, and David C Burr (1998). "Seeing biological motion". In: *Nature* 395.6705, pp. 894–896.

Neville, Helen J, Daphne Bavelier, David Corina, Josef Rauschecker, Avi Karni, Anil Lalwani, Allen Braun, Vince Clark, Peter Jezzard, and Robert Turner (1998). "Cerebral organization for language in deaf and hearing subjects: biological constraints and effects of experience". In: *Proceedings of the National Academy of Sciences* 95.3, pp. 922–929.

Newman, Aaron J, Daphne Bavelier, David Corina, Peter Jezzard, and Helen J Neville (2002). "A critical period for right hemisphere recruitment in American Sign Language processing". In: *Nature neuroscience* 5.1, pp. 76–80.

Niyogi, Sourabh A, Edward H Adelson, et al. (1994). "Analyzing and recognizing walking figures in XYT". In: *CVPR*. Vol. 94, pp. 469–474.

Norman-Haignere, Sam V and Josh H McDermott (2018). "Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex". In: *PLoS biology* 16.12, e2005127.

Nyquist, Harry (1928). "Certain topics in telegraph transmission theory". In: *Transactions of the American Institute of Electrical Engineers* 47.2, pp. 617–644.

O'Toole, Alice J, Hervé Abdi, Kenneth A Deffenbacher, and Dominique Valentin (1993). "Low-dimensional representation of faces in higher dimensions of the face space". In: *JOSA A* 10.3, pp. 405–411.

Okada, Kayoko, Corianne Rogalsky, Lucinda O'Grady, Leila Hanaumi, Ursula Bellugi, David Corina, and Gregory Hickok (2016). "An fMRI study of perception and action in deaf signers". In: *Neuropsychologia* 82, pp. 179–188.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu

(2016). "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499*.

OptiTrack NaturalPoint, Inc. *Motive 2.1 | What's New*. URL: https://youtu.be/KRvQhYJtIUI.

— *Optitrack*. URL: https://optitrack.com/.

Oszust, Mariusz and Marian Wysocki (2013). "Polish sign language words recognition with Kinect". In: *2013 6th International Conference on Human System Interactions (HSI)*. IEEE, pp. 219–226.

Oz, Cemil and Ming C Leu (2011). "American sign language word recognition with a sensory glove using artificial neural networks". In: *Engineering Applications of Artificial Intelligence* 24.7, pp. 1204–1213.

Pariente, Manuel, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent (2020). "Filterbank Design for End-to-end Speech Separation". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6364–6368. DOI: 10.1109/ICASSP40776.2020.9053038.

PhaseSpace. URL: https://phasespace.com/.

Poizner, Howard, Ursula Bellugi, and Venita Lutes-Driscoll (1981). "Perception of American sign language in dynamic point-light displays." In: *Journal of experimental psychology: Human perception and performance* 7.2, p. 430.

Poizner, Howard, Edward S Klima, Ursula Bellugi, and Robert B Livingston (1986). "Motion analysis of grammatical processes in a visual-gestural language". In: *Event cognition: An ecological perspective*, pp. 155–174.

Portilla, Javier and Eero P Simoncelli (2000). "A parametric texture model based on joint statistics of complex wavelet coefficients". In: *International journal of computer vision* 40.1, pp. 49–70.

Prinz, Wolfgang (1997). "Perception and action planning". In: *European journal of cognitive psychology* 9.2, pp. 129–154.

Promsri, Arunee and Peter Federolf (2020). "Analysis of postural control using principal component analysis: The relevance of postural accelerations and of their frequency dependency for selecting the number of movement components". In: *Frontiers in Bioengineering and Biotechnology* 8, p. 480.

Pu, Junfu, Wengang Zhou, and Houqiang Li (2016). "Sign language recognition with multi-modal features". In: *Pacific Rim Conference on Multimedia*. Springer, pp. 252–261.

Rasamimanana, Nicolas, Frederic Bevilacqua, Norbert Schnell, Fabrice Guedy, Emmanuel Flety, Come Maestracci, Bruno Zamborlin, Jean-Louis Frechin, and Uros Petrevski (2010). "Modular musical objects towards embodied control of digital music". In: *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*, pp. 9–12.

Reed, Catherine L and Martha J Farah (1995). "The psychological reality of the body schema: a test with normal participants." In: *Journal of Experimental Psychology: Human Perception and Performance* 21.2, p. 334.

Reid, Robert C (2010). "A kinematic and kinetic study of alpine skiing technique in slalom". In:

Rizzolatti, Giacomo and Laila Craighero (2004). "The mirror-neuron system". In: *Annu. Rev. Neurosci.* 27, pp. 169–192.

Rizzolatti, Giacomo, Leonardo Fogassi, and Vittorio Gallese (2001). "Neurophysiological mechanisms underlying the understanding and imitation of action". In: *Nature reviews neuroscience* 2.9, pp. 661–670.

Rousselet, Guillaume A, Marc J-M Macé, and Michèle Fabre-Thorpe (2003). "Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes". In: *Journal of vision* 3.6, pp. 5–5.

Runeson, Sverker and Gunilla Frykholm (1981). "Visual perception of lifted weight." In: *Journal of Experimental Psychology: Human Perception and Performance* 7.4, p. 733.

— (1983). "Kinematic specification of dynamics as an informational basis for person-and-action perception: expectation, gender recognition, and deceptive intention." In: *Journal of experimental psychology: general* 112.4, p. 585.

Saggio, Giovanni, Pietro Cavallo, Mariachiara Ricci, Vito Errico, Jonathan Zea, and Marco E Benalcázar (2020). "Sign language recognition using wearable electronics: implementing k-nearest neighbors with dynamic time warping and convolutional neural network algorithms". In: *Sensors* 20.14, p. 3879.

Sallandre, Marie-Anne and Christian Cuxac (2002). "Iconicity in Sign Language: a theoretical and methodological point of view". In: *Lecture notes in computer science*, pp. 173–180.

Saltzman, Elliot (1979). "Levels of sensorimotor representation". In: *journal of Mathematical Psychology* 20.2, pp. 91–163.

Sarasúa, Álvaro and Enric Guaus (2014). "Dynamics in Music Conducting: A Computational Comparative Study Among Subjects." In: *NIME*, pp. 195–200.

Sartori, Luisa, Andrea Camperio, Maria Bulgheroni, and Umberto Castiello (2013). "Reach-to-grasp movements in Macaca fascicularis monkeys: the Isochrony Principle at work". In: *Frontiers in psychology* 4, p. 114.

Schoonderwaldt, Erwin, Nicolas H Rasamimanana, and Frédéric Bevilacqua (2006). "Combining accelerometer and video camera: Reconstruction of bow velocity profiles". In: *6th International Conference on New Interfaces for Musical Expression*, pp. 1–1.

Schwab, Arend L and Jaap P Meijaard (2006). "How to draw Euler angles and utilize Euler parameters". In: *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Vol. 42568. Citeseer, pp. 259–265.

Segouat, Jérémie and Annelies Braffort (2009). "Toward the study of sign language coarticulation: methodology proposal". In: *2009 Second International Conferences on Advances in Computer-Human Interactions*. IEEE, pp. 369–374.

Sevdalis, Vassilis and Peter E Keller (2009). "Self-recognition in the Perception of Actions Performed in Synchrony with Music". In: *Annals of the New York Academy of Sciences* 1169.1, pp. 499–502.

Shannon, Claude Elwood (1949). "Communication in the presence of noise". In: *Proceedings of the IRE* 37.1, pp. 10–21.

Shi, Bowen, Aurora Martinez Del Rio, Jonathan Keane, Jonathan Michaux, Diane Brentari, Greg Shakhnarovich, and Karen Livescu (2018). "American sign language fingerspelling recognition in the wild". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 145–152.

Sie, Mu-Syuan, Yu-Chuan Cheng, and Cheng-Chin Chiang (2014). "Key motion spotting in continuous motion sequences using motion sensing devices". In: *2014 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. IEEE, pp. 326–331.

Sigal, Leonid, David J Fleet, Nikolaus F Troje, and Micha Livne (2010). "Human attributes from 3d pose tracking". In: *European conference on computer vision*. Springer, pp. 243–257.

Simhi, Noa and Galit Yovel (2020). "Dissociating gait from static appearance: A virtual reality study of the role of dynamic identity signatures in person recognition". In: *Cognition* 205, p. 104445.

Sirovich, Lawrence and Michael Kirby (1987). "Low-dimensional procedure for the characterization of human faces". In: *Josa a* 4.3, pp. 519–524.

Skogstad, Ståle Andreas van Dorp, Kristian Nymoen, Mats Erling Høvin, Sverre Holm, and Alexander Refsum Jensenius (2013). "Filtering motion capture data for real-time applications". In:

Sperling, George, Michael Landy, Yoav Cohen, and M Pavel (1985). "Intelligible encoding of ASL image sequences at extremely low information rates". In: *Computer vision, graphics, and image processing* 31.3, pp. 335–391.

Starner, Thad and Alex Pentland (1997). "Real-time american sign language recognition from video using hidden markov models". In: *Motion-based recognition*. Springer, pp. 227–243.

Stevenage, Sarah V, Mark S Nixon, and Kate Vince (1999). "Visual analysis of gait as a cue to identity". In: *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 13.6, pp. 513–526.

Stoll, Chloé, Richard Palluel-Germain, Roberto Caldara, Junpeng Lao, Matthew WG Dye, Florent Aptel, and Olivier Pascalis (2018). "Face recognition is shaped by the use of sign language". In: *The Journal of Deaf Studies and Deaf Education* 23.1, pp. 62–70.

Sutton, Valerie (2009). "SignWriting". In: *Sl: sn*, p. 9.

Systems, Vicon Motion. *Vicon*. URL: https://vicon.com/.

Tamura, Shinichi and Shingo Kawasaki (1988). "Recognition of sign language motion images". In: *Pattern recognition* 21.4, pp. 343–353.

Tartter, Vivien C and Kenneth C Knowlton (1981). "Perception of sign language from an array of 27 moving spots". In: *Nature* 289.5799, pp. 676–678.

Tennant, Richard A, Marianne Gluszak, and Marianne Gluszak Brown (1998). *The American sign language handshape dictionary*. Gallaudet University Press.

Thelen, Esther and Donna M Fisher (1983). "The organization of spontaneous leg movements in newborn infants". In: *Journal of motor behavior* 15.4, pp. 353–372.

Tilmanne, Joëlle, Nicolas d'Alessandro, Maria Astrinaki, and Thierry Ravet (2014). "Exploration of a stylistic motion space through realtime synthesis". In: *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*. Vol. 2. IEEE, pp. 803–809.

Tilmanne, Joëlle and Thierry Dutoit (2010). "Expressive gait synthesis using PCA and Gaussian modeling". In: *International Conference on Motion in Games*. Springer, pp. 363–374.

Tilmanne, Joëlle, Alexis Moinet, and Thierry Dutoit (2012). "Stylistic gait synthesis based on hidden Markov models". In: *EURASIP Journal on advances in signal processing* 2012.1, pp. 1–14.

Tits, Mickaël (2018). "Expert Gesture Analysis through Motion Capture using Statistical Modeling and Machine Learning". PhD thesis. Ph. D. Dissertation.

Tits, Mickaël, Joëlle Tilmanne, Nicolas D'Alessandro, and Marcelo M Wanderley (2015). "Feature extraction and expertise analysis of pianists' Motion-Captured Finger Gestures". In: *ICMC*.

Tits, Mickaël, Joëlle Tilmanne, and Thierry Dutoit (2017). "Morphology independent feature engineering in motion capture database for gesture evaluation". In: *Proceedings of the 4th International Conference on Movement Computing*, pp. 1–8.

— (2018). "Robust and automatic motion-capture data recovery using soft skeleton constraints and model averaging". In: *PloS one* 13.7, e0199744.

Todorov, Emanuel and Zoubin Ghahramani (2004). "Analysis of the synergies underlying complex hand manipulation". In: *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 2. IEEE, pp. 4637–4640.

Toiviainen, Petri, Geoff Luck, and Marc R Thompson (2010). "Embodied meter: hierarchical eigenmodes in music-induced movement". In: *Music Perception* 28.1, pp. 59–70.

Tolba, AS, AH El-Baz, and AA El-Harby (2006). "Face recognition: A literature review". In: *International Journal of Signal Processing* 2.2, pp. 88–103.

Troje, Nikolaus F (2002a). "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns". In: *Journal of vision* 2.5, pp. 2–2.

— (2002b). "The little difference: Fourier based synthesis of gender-specific biological motion". In: *Dynamic perception*, pp. 115–120.

Troje, Nikolaus F and Thomas Vetter (1996). "Representations of human faces". In:

Troje, Nikolaus F, Cord Westhoff, and Mikhail Lavrov (2005). "Person identification from biological motion: Effects of structural and kinematic cues". In: *Perception & Psychophysics* 67.4, pp. 667–675.

Tyrone, Martha E, Hosung Nam, Elliot Saltzman, Gaurav Mathur, and Louis Goldstein (2010). "Prosody and movement in American Sign Language: A task-dynamics approach". In: *Speech Prosody 2010-Fifth International Conference*.

Uebersax, Dominique, Juergen Gall, Michael Van den Bergh, and Luc Van Gool (2011). "Real-time sign language letter and word recognition from depth data". In: *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*. IEEE, pp. 383–390.

Unuma, Munetoshi and Ryozo Takeuchi (1991). "Generation of human motion with emotion". In: *Computer Animation'91*. Springer, pp. 77–88.

Valentin, Dominique, Hervé Abdi, Alice J O'Toole, and Garrison W Cottrell (1994). "Connectionist models of face processing: A survey". In: *Pattern recognition* 27.9, pp. 1209–1230.

Veale, Tony, Alan Conway, and Bróna Collins (1998). "The challenges of cross-modal translation: English-to-Sign-Language translation in the Zardoz system". In: *Machine Translation* 13.1, pp. 81–106.

Venture, Gentiane, Hideki Kadone, Tianxiang Zhang, Julie Grèzes, Alain Berthoz, and Halim Hicheur (2014). "Recognizing emotions conveyed by human gait". In: *International Journal of Social Robotics* 6.4, pp. 621–632.

Véray, Laurent (2007). *« Le sport et la photographie scientifique », Histoire par l'image*. URL: http://histoire-image.org/fr/etudes/sport-photographie-scientifique.

Vetter, Thomas and Nikolaus F Troje (1995). "A separate linear shape and texture space for modeling two-dimensional images of human faces". In:

— (1997). "Separation of texture and shape in images of faces for image coding and synthesis". In: *JOSA A* 14.9, pp. 2152–2161.

Vinson, David P, Kearsy Cormier, Tanya Denmark, Adam Schembri, and Gabriella Vigliocco (2008). "The British Sign Language (BSL) norms for age of acquisition, familiarity, and iconicity". In: *Behavior research methods* 40.4, pp. 1079–1087.

Vinter, Annie and Pierre Mounoud (1991). "Isochrony and accuracy of drawing movements in children: Effects of age and context". In: *Development of Graphic Skills. Research Perspectives and Educational Implications*, pp. 113–134.

Viviani, Paolo and Gin McCollum (1983). "The relation between linear extent and velocity in drawing movements". In: *Neuroscience* 10.1, pp. 211–218.

Viviani, Paolo and Roland Schneider (1991). "A developmental study of the relationship between geometry and kinematics in drawing movements." In: *Journal of Experimental Psychology: Human Perception and Performance* 17.1, p. 198.

Viviani, Paolo and Natale Stucchi (1992). "Biological movements look uniform: evidence of motor-perceptual interactions." In: *Journal of experimental psychology: Human perception and performance* 18.3, p. 603.

Von Agris, Ulrich and Karl-Friedrich Kraiss (2007). "Towards a video corpus for signer-independent continuous sign language recognition". In: *Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, May* 11.

Watanabe, Katsumi, Tetsuya Matsuda, Tomoyuki Nishioka, and Miki Namatame (2011). "Eye gaze during observation of static faces in deaf people". In: *PloS one* 6.2, e16919.

Weast, Traci Patricia (2008). *Questions in American Sign Language: A quantitative analysis of raised and lowered eyebrows*. The University of Texas at Arlington.

Welch, Peter (1967). "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms". In: *IEEE Transactions on audio and electroacoustics* 15.2, pp. 70–73.

Westhoff, Cord and Nikolaus F Troje (2007). "Kinematic cues for person identification from biological motion". In: *Perception & psychophysics* 69.2, pp. 241–253.

WFD, World Federation of the Deaf (2016). *"Our work"*. URL: http://wfdeaf.org/our-work/.

Wilbur, Ronnie B (1999). "Stress in A SL: Empirical Evidence and Linguistic I ssues". In: *Language and speech* 42.2-3, pp. 229–250.

Wilbur, Ronnie B and Brenda S Schick (1987). "The effects of linguistic stress on ASL signs". In: *Language and Speech* 30.4, pp. 301–323.

Windows, Kinect for. *Kinect*. URL: https://developer.microsoft.com/en-us/windows/kinect/.

Winter, David A (2009). *Biomechanics and motor control of human movement*. John Wiley & Sons.

Wolfe, Rosalee, John McDonald, and Jerry C Schnepp (2011). "Avatar to depict sign language: Building from reusable hand animation". In:

Wolpert, Daniel M and Zoubin Ghahramani (2000). "Computational principles of movement neuroscience". In: *Nature neuroscience* 3.11, pp. 1212–1217.

Xsens (2021). *A history of motion capture*. URL: https://www.xsens.com/a-history-of-motion-capture.

Xsens gloves. URL: https://www.xsens.com/products/xsens-gloves-by-manus.

Xsens Mtw Awinda. URL: https://www.xsens.com/products/mtw-awinda.

Yan, Yichao, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang (2016). "Person re-identification via recurrent feature aggregation". In: *European Conference on Computer Vision*. Springer, pp. 701–716.

Yan, Yuke, James M Goodman, Dalton D Moore, Sara A Solla, and Sliman J Bensmaia (2020). "Unexpected complexity of everyday manual behaviors". In: *Nature communications* 11.1, pp. 1–8.

Young, Cole and David J Reinkensmeyer (2014). "Judging complex movement performances for excellence: a principal components analysis-based technique applied to competitive diving". In: *Human Movement Science* 36, pp. 107–122.

Zafrulla, Zahoor, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti (2011). "American sign language recognition with the kinect". In: *Proceedings of the 13th international conference on multimodal interfaces*, pp. 279–286.

Zago, Matteo, Marina Codari, F Marcello Iaia, and Chiarella Sforza (2017a). "Multi-segmental movements as a function of experience in karate". In: *Journal of sports sciences* 35.15, pp. 1515–1522.

Zago, Matteo, Ilaria Pacifici, Nicola Lovecchio, Manuela Galli, Peter Andreas Federolf, and Chiarella Sforza (2017b). "Multi-segmental movement patterns reflect juggling complexity and skill level". In: *Human movement science* 54, pp. 144–153.

Zago, Matteo, Chiarella Sforza, Alessia Bona, Veronica Cimolin, Pier Francesco Costici, Claudia Condoluci, and Manuela Galli (2017c). "How multi segmental patterns deviate in spastic diplegia from typical developed". In: *Clinical Biomechanics* 48, pp. 103–109.

Zahedi, Morteza, Philippe Dreuw, David Rybach, Thomas Deselaers, and Hermann Ney (2006). "Continuous sign language recognition-approaches from speech recognition and available data resources". In: *Second Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios*, pp. 21–24.

Zahedi, Morteza, Daniel Keysers, Thomas Deselaers, and Hermann Ney (2005). "Combination of tangent distance and an image distortion model for appearance-based sign language recognition". In: *Joint Pattern Recognition Symposium*. Springer, pp. 401–408.

Zhang, Zonghua and Nikolaus F Troje (2005). "View-independent person identification from human gait". In: *Neurocomputing* 69.1-3, pp. 250–256.

Zhao, Liwei, Karin Kipper, William Schuler, Christian Vogler, Norman Badler, and Martha Palmer (2000). "A machine translation system from English to American Sign Language". In: *Conference of the Association for Machine Translation in the Americas*. Springer, pp. 54–67.

Appendix A

Description of the pictures used in the MOCAP1 corpus

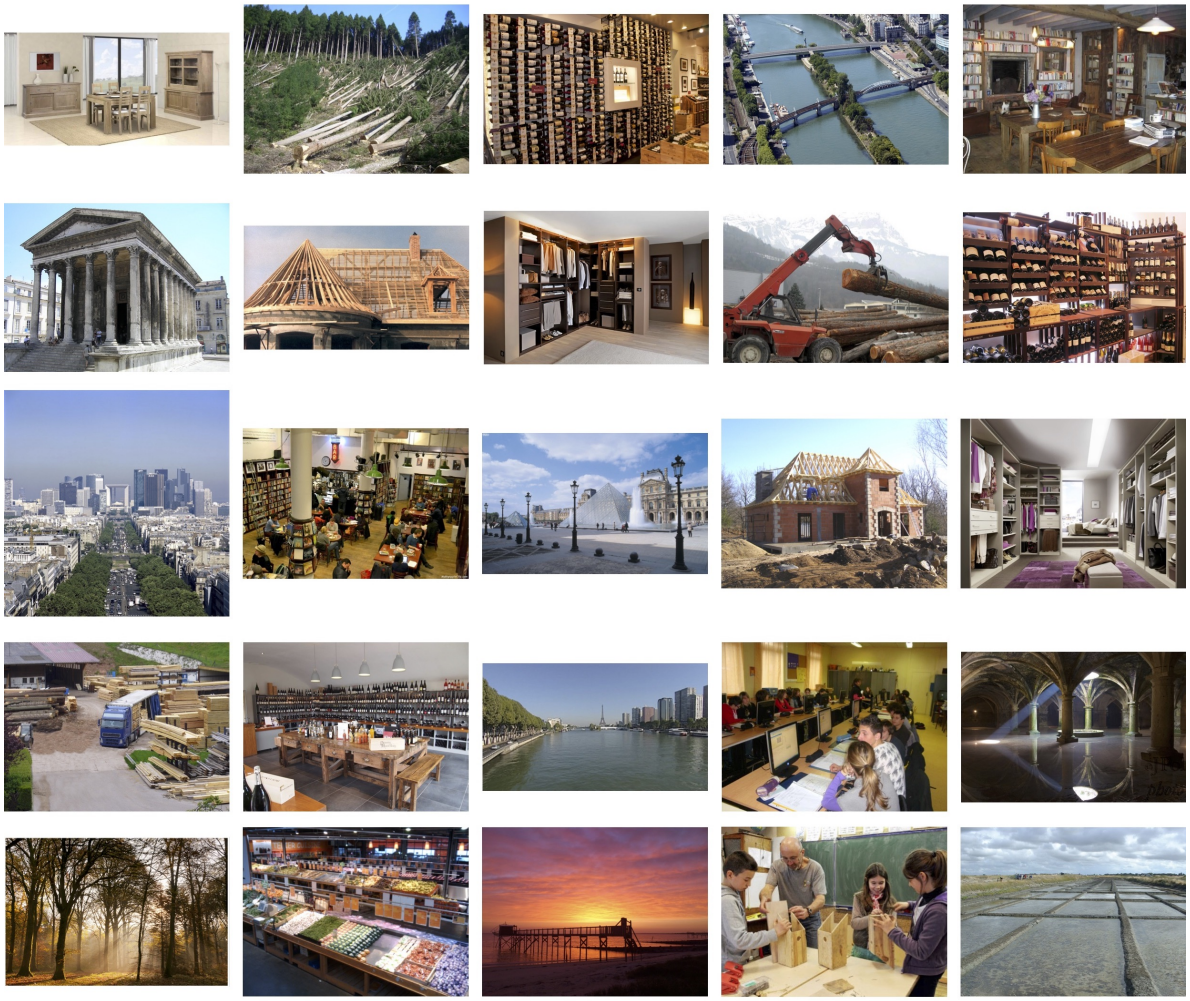


FIGURE A.1: The 25 pictures described by the signers in MOCAP1 corpus (Benchiheub et al., 2016b; Benchiheub et al., 2016a).

Appendix B

Correspondences between labels of MOCAP1 and MOCAP1-v2

TABLE B.1: Correspondences between signer labels of MOCAP1 original corpus and the new version MOCAP1-v2 used in the present thesis. *not present in the public release (Benchiheub et al., 2016a).

MOCAP1-v2	MOCAP1
Signer 1	l2
Signer 2	l6*
Signer 3	l8*
Signer 4	l4
Signer 5	l1
Signer 6	l3

TABLE B.2: Correspondences between mocap example labels of MOCAP1 original corpus and the new version MOCAP1-v2 used in the present thesis. Example labels correspond to the pictures described by signers.

MOCAP1-v2	MOCAP1
IM01	IM01
IM02	IM02
IM03	IM03
IM04	IM04
IM05	IM05
IM06	IM06
IM07	IM08
IM08	IM09
IM09	IM10
IM10	IM11
IM11	IM12
IM12	IM13
IM13	IM14
IM14	IM15
IM15	IM16
IM16	IM17
IM17	IM18
IM18	IM19
IM19	IM20
IM20	IM21
M21	IM22
IM22	IM23
IM23	IM24
IM24	IM25

Appendix C

The role of the statistics in the automatic signer identification

TABLE C.1: Bonferroni-adjusted post Hoc. Comparisons of the identification accuracy of the model using different subsets of statistics.

		Mean Difference	SE	t	p
stdpos	+ skewpos	0.042	0.044	0.951	1.000
f	+ kurtpos	0.007	0.044	0.158	1.000
	+ meanvel	0.021	0.044	0.475	1.000
	+ stdvel	-0.132	0.044	-3.019	0.083
	+ skewvel	-0.083	0.044	-1.907	1.000
	+ kurtvel	-0.111	0.044	-2.543	0.334
	+ cov_vel	-0.201	0.044	-4.607	< .001***
+ skewpos	+ kurtpos	-0.035	0.044	-0.793	1.000
	+ meanvel	-0.021	0.044	-0.476	1.000
	+ stdvel	-0.174	0.044	-3.970	0.003**
	+ skewvel	-0.125	0.044	-2.858	0.135
	+ kurtvel	-0.153	0.044	-3.494	0.017*
	+ cov_vel	-0.243	0.044	-5.558	< .001***
+ kurtpos	+ meanvel	0.014	0.044	0.316	1.000
	+ stdvel	-0.139	0.044	-3.177	0.050*
	+ skewvel	-0.090	0.044	-2.066	1.000
	+ kurtvel	-0.118	0.044	-2.701	0.214
	+ cov_vel	-0.208	0.044	-4.766	< .001***
+ meanvel	+ stdvel	-0.153	0.044	-3.494	0.017
	+ skewvel	-0.104	0.044	-2.382	0.515
	+ kurtvel	-0.132	0.044	-3.017	0.083
	+ cov_vel	-0.222	0.044	-5.082	< .001***
+ stdvel	+ skewvel	0.049	0.044	1.112	1.000
	+ kurtvel	0.021	0.044	0.476	1.000
	+ cov_vel	-0.069	0.044	-1.588	1.000
+ skewvel	+ kurtvel	-0.028	0.044	-0.635	1.000
	+ cov_vel	-0.118	0.044	-2.700	0.215
+ kurtvel	+ cov_vel	-0.090	0.044	-2.065	1.000

* $p < .05$, ** $p < .01$, *** $p < .001$

Appendix D

Synthesis example 3: identity conversion from Signer 2 to Signer 1

In this third example, synthesized mocap excerpts were generated in order to make the movements of Signer 2 identified as those of Signer 1. For that aim, the synthesis procedure was driven by the following target statistics:

$$\tilde{\boldsymbol{d}}_{\mathbf{50}} = \boldsymbol{d}_{orig} + 50\boldsymbol{d}_{\mathbf{1}} \tag{D.1}$$

where $\boldsymbol{d}_{orig}$ are the statistics of the mocap example 1 of Signer 2, and $\boldsymbol{d}_{\mathbf{1}}$ are the discriminant statistical patterns of Signer 1.

The synthesis procedure involved 2,000 iterations and a step size of 0.001. The PLD videos of the original and synthesized mocap excerpts can be found in Videos 10.7 and 10.8. The successful modifications of the synthesis procedure were confirmed by the automatic signer identification model, which identified the synthesized motion as that of Signer 1 while it identified the original motion as produced by Signer 2:

TABLE D.1: Output of the automatic signer identification model. P is the probability that the movements were produced by the signer (see Equation 9.4).

Original mocap example	Synthesized mocap example
$P(S = 2) = 0.99$	$P(S = 1) = 0.99$

Further visualizations of the statistical modifications (Figure D.1) and how it affected the mocap signals (Figure D.2) are available below.

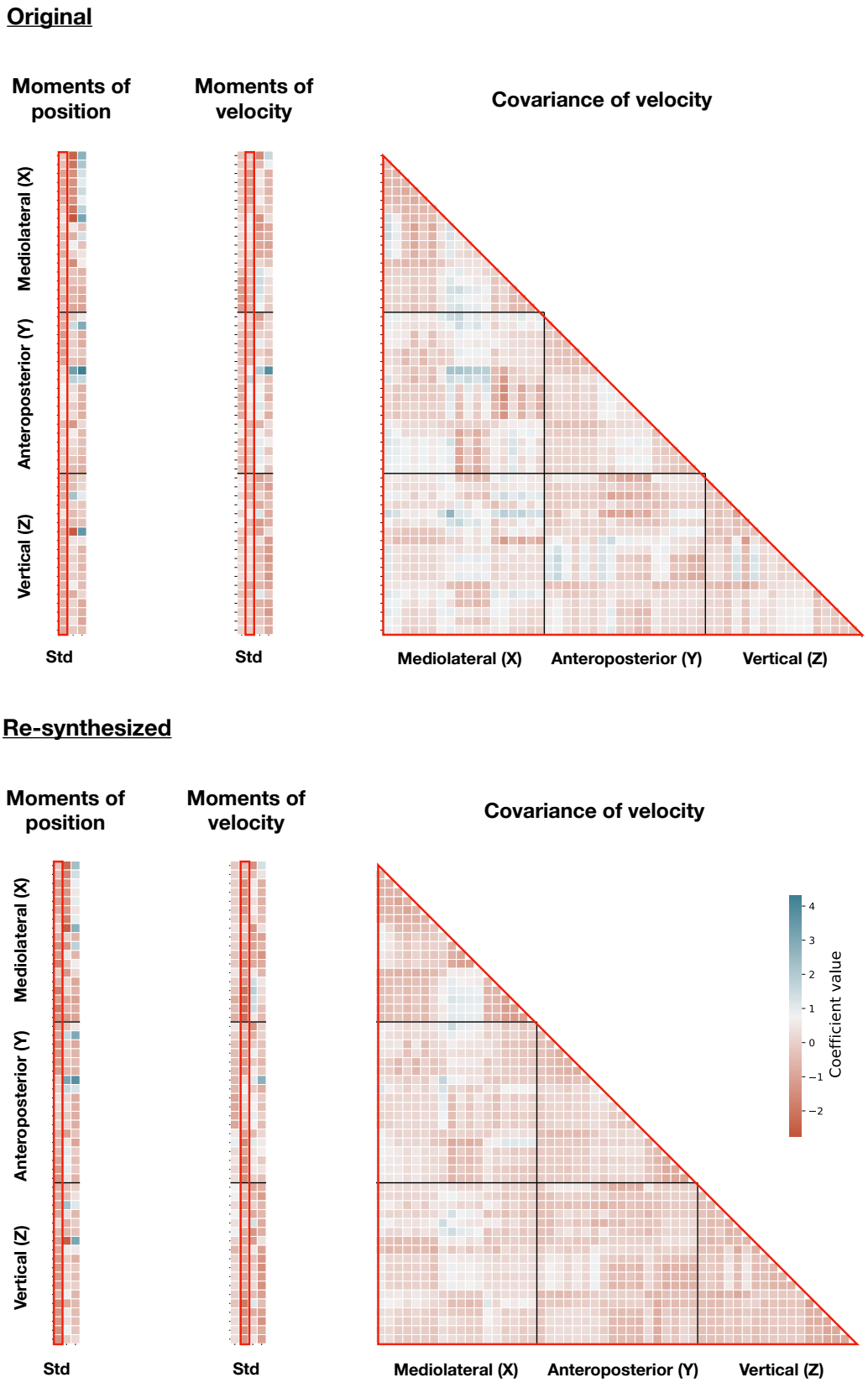


FIGURE D.1: Statistics of kinematic features of mocap example 1 of Signer 2, before (original) and after (re-synthesized) the synthesis procedure: moments (columns: std, skew, kurtosis) of position for all markers (rows), moments (columns: mean, std, skew, kurtosis) of velocity for all markers (rows), covariance of velocity between markers (rows and columns). Markers are sorted from 1 to 19 as presented in Section 5.3.1 along X, Y and Z axes. Blue represents positive statistical values, while red represents negative ones. The three main classes of statistics that carry identity information are highlighted.

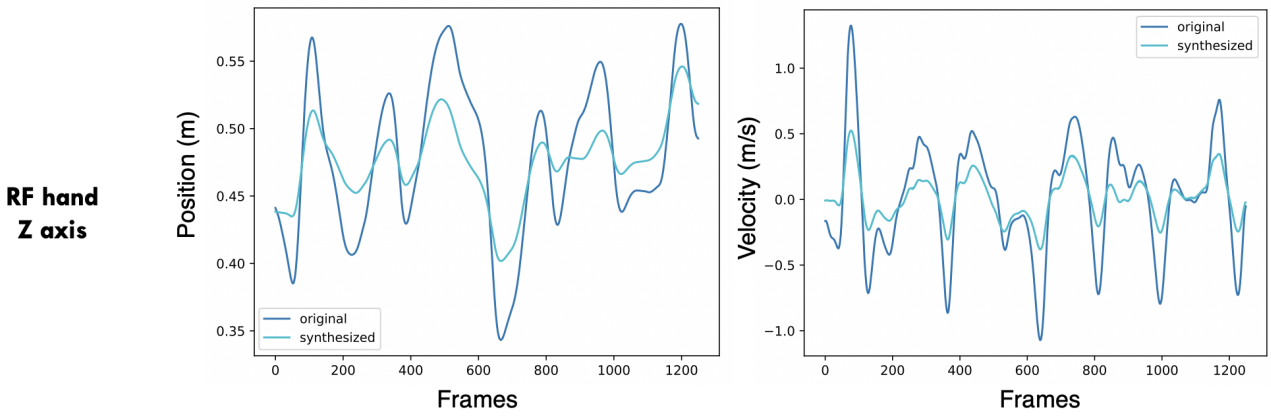


FIGURE D.2: Position and velocity data of the RF hand marker along the Z axis, for the original and synthesized movements.

Appendix E

Publications and communications during the PhD

Peer-reviewed journal

- Bigand, F., Prigent, E., Berret, B., & Braffort, A. (2021). Decomposing spontaneous sign language into elementary movements: A principal component analysis-based approach. PloS one, 16(10), e0259464.
- Bigand, F., Prigent, E., Berret, B., & Braffort, A. (2021). Machine Learning of Motion Statistics Reveals the Kinematic Signature of the Identity of a Person in Sign Language. Frontiers in Bioengineering and Biotechnology, 603.

Peer-reviewed conference, with proceedings

- Bigand, F., Prigent, E., Berret, B., & Braffort, A. (2021, August). How Fast is Sign Language? A Reevaluation of the Kinematic Bandwidth using Motion Capture. To be published in Proceedings of the 29th EUSIPCO.
- Bigand, F., Prigent, E., & Braffort, A. (2020, July). Person Identification Based On Sign Language Motion: Insights From Human Perception and Computational Modeling. In Proceedings of the 7th International Conference on Movement and Computing (pp. 1-7).
- Bigand, F., Prigent, E., & Braffort, A. (2019, October). Retrieving Human Traits from Gesture in Sign Language: The Example of Gestural Identity. In Proceedings of the 6th International Conference on Movement and Computing (pp. 1-4).
- Bigand, F., Prigent, E., & Braffort, A. (2019, July). Animating virtual signers: the issue of gestural anonymization. In Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (pp. 252-255).

Peer-reviewed conference, without proceedings

- Bigand, F., Braffort, A., Prigent, E., & Berret, B. (2019, September). Signing Avatar Motion: Combining Naturality and Anonymity. In International Workshop on Sign Language Translation and Avatar Technology.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Extraction de caractéristiques humaines dans le mouvement par apprentissage automatique : l'exemple de l'identité en Langue des Signes

Mots clés : Mouvement, Machine Learning, Extraction de caractéristiques, Capture de mouvement, Langue des Signes, Perception

Résumé : De nombreux obstacles technologiques doivent être surmontés afin d'outiller les Langues des Signes (LS) de la même manière que les langues parlées. Pour ce faire, il est nécessaire d'approfondir les connaissances dans de multiples disciplines, en particulier les sciences du mouvement. Plus précisément, cette thèse vise à étudier la possibilité d'anonymiser les mouvements d'un signeur, de la même manière qu'un locuteur peut rester anonyme en modifiant des aspects spécifiques de sa voix.

Premièrement, cette thèse met en lumière les propriétés cinématiques de la LS spontanée afin d'améliorer les modèles de LS naturelle. En utilisant des données de mouvements en 3D de plusieurs signeurs, nous montrons que la bande passante cinématique de la LS spontanée diffère fortement de celle des signes isolés. Ensuite, une analyse en composantes principales révèle que les discours spontanés peuvent être décrits par un ensemble réduit de mouvements simples (i.e., synergies).

De plus, en combinant données humaines et modélisation informatique, cette thèse démontre que les signeurs peuvent être identifiés à partir de leurs mouvements, au-delà de la morphologie et de la posture. Enfin, nous présentons des modèles d'apprentissage automatique capables d'extraire automatiquement l'information d'identité dans les mouvements de la LS, puis de la manipuler lors de la génération. Les modèles développés dans cette thèse pourraient permettre de produire des messages de LS anonymes via des signeurs virtuels, ce qui ouvrirait de nouveaux horizons aux signeurs sourds.

Title: Extracting human characteristics from motion using machine learning: the case of identity in Sign Language

Keywords: Motion, Machine Learning, Feature extraction, Motion capture, Sign Language, Perception

Abstract: Many technological barriers must be tackled in order to provide tools in Sign Languages (SLs) in the same way as for spoken languages. For that aim, further insights must be gained into multiple disciplines, in particular motion science. More specifically, the present thesis aims to gain insights into the possibility of anonymizing the movements of a signer, in the same way as a speaker can remain anonymous by modifying specific aspects of the voice.

First, this thesis sheds light on general kinematic properties of spontaneous SL in order to improve the models of natural SL. Using 3D motion recordings of multiple signers, we show that the kinematic bandwidth of spontaneous SL highly differs from that of signs made in isolation. Furthermore, a Principal Component Analysis reveals that the spontaneous SL discourses can be described by a reduced set of simple, one-directional, movements (i.e., synergies).

Furthermore, combining human data and computational modelling, we demonstrate that signers can be identified from their movements, beyond morphology- and posture-related cues. Finally, we present machine learning models able to automatically extract identity information in SL movements and to manipulate it in generated motion. The models developed in this thesis could allow producing anonymized SL messages via virtual signers, which would open new horizons for deaf SL users.