

Human movement representation on multivariate time series for recognition of professional gestures and forecasting their trajectories

Sotiris Manitsaris^{1*}, Gavriela Senteri¹, Dimitrios Makrygiannis¹, Alina Glushkova¹,

¹Centre for Robotics, MINES ParisTech, PSL Université Paris, France

* Correspondence:

Sotiris Manitsaris

sotiris.manitsaris@mines-paristech.

Keywords: State Space representation, differential equations, movement modeling, Hidden Markov Models, gesture recognition, forecasting, motion trajectory

Abstract

Human-Centered Artificial Intelligence is increasingly deployed in professional workplaces in Industry 4.0 to address various challenges related to the collaboration between the operators and the machines, the augmentation of their capabilities or the improvement of the quality of their work and life in general. Intelligent systems and autonomous machines need to continuously recognize and follow the professional actions and gestures of the operators in order to collaborate with them and anticipate their trajectories for avoiding potential collisions and accidents. Nevertheless, the recognition of patterns of professional gestures is a very challenging task for both research and the industry. There are various types of human movements that the intelligent systems need to perceive, e.g., gestural commands to machines, professional actions with or without the use of tools etc. Moreover, the *inter-* and *intra-* class spatiotemporal variances together with the very limited access to annotated human motion data constitute a major research challenge. In this paper, we introduce the Gesture Operational Model which describes how gestures are performed based on assumptions that focus on the dynamic association of body entities, their synergies, and their serial and non-serial mediations, as well as, their transitioning over time from one state to another. Then, the assumptions of the Gesture Operational Model are translated into a simultaneous equations system for each body entity through State-Space modeling. The coefficients of the equation are computed using the Maximum Likelihood Estimation method. The simulation of the model generates a confidence-bounding box for every entity that describes the tolerance of its spatial variance over time. The contribution of our approach is demonstrated both for recognizing gestures and forecasting human motion trajectories. In recognition, it is combined with continuous Hidden Markov Models to boost the recognition accuracy when the likelihoods are not confident. In forecasting, a motion trajectory can be estimated by taking as minimum input two observations only. The performance of the algorithm has been evaluated using four industrial datasets that contain gestures and actions from a TV assembly line, the glassblowing industry, the gestural commands to Automated Guided Vehicles as well as the Human-Robot Collaboration in the automotive assembly lines.

1 Introduction

Human motion analysis and recognition is widely researched from various scientific domains including Human-Computer Interaction, Collaborative Robotics and Autonomous Vehicles. Both the industry

and science, face significant challenges in capturing the human motion, developing models and algorithms for efficiently recognizing it, as well as for improving the perception of the machines when collaborating with humans.

Nevertheless, in factories, « *we always start with manual work* » as explained by Mitsuri Kawai, Head of Manufacturing and Executive Vice-President of Toyota [50]. Therefore, experts from both collaborative robotics and applied ergonomics are always involved when a new collaborative cell is being designed. Nowadays, despite the significant progress in training robots by demonstration, automatizing the human tasks in mixed workspaces, still remains the goal. However, those workspaces are not necessarily collaborative. For example, in a smart workspace, a machine that perceives and anticipates gestures and actions of the operator, would be able to adapt its own actions depending on those of the operator, thus giving him/her the possibility to obtain ergonomically ‘green postures’. Furthermore, Automated Guided Vehicles (AGVs) should also be able to detect the intentions of the operator with the aim of collaborating with them, avoiding accidents and understanding gestural commands. Finally, in Industry 4.0, an important number of Creative and Cultural Industries, e.g. in luxury goods manufacturing, still base their know-how on manual dexterity, no matter whether the operator is in collaboration with a machine or not. Therefore, human movement representation and gesture recognition constitute a mean for identifying the industrial know-how and transmitting it to the next generation of the operators.

From a scientific point of view, major research challenges are faced by scientists, especially when dealing with professional environments in an industrial context. Initially, there is an extremely limited access to motion data from real-life configurations. This is mainly due to acceptability issues from the operators or to limitations imposed by laws and regulations that protect the access to/use of personal data, e.g. the ‘General Data Protection Regulation’ in the European Union. Therefore, most of existing datasets include only gestures from the everyday life. Furthermore, when creating custom datasets with professional motion data, many practical and environmental issues might occur, e.g. variation in luminosity, various workspace with different geometries, camera in motion to record a person moving in space, low availability of real experts etc. Additionally, the community of actions and gesture recognition deals with challenges that are related to intra- and *inter-* class variations [49]. Frequent are the cases where a professional task involves gestures that have very similar spatiotemporal characteristics - *low inter- class variation* – together with very important differences in the way different humans perform the same gesture -*high intra- class variation-*. Finally, when applied to accident prevention, a small delay in predicting the action might be crucial.

The work presented in this paper, contributes to the aforementioned challenges, through the proposition of a Gesture Operational Model (GOM) that describes how the body parts cooperate, to perform a situated professional gesture. The model is built upon several assumptions that determine the dynamic relationship between the body entities within the execution of the human movement. The model is based on the State-Space (SS) representation, as a simultaneous equation system for all the body entities is generated, composed by a set of first-order differential equations. The coefficients of the equation system are estimated using the Maximum Likelihood Estimation method (MLE), and its dynamic simulation generates a dynamic tolerance of the spatial variance of the movement over time. The scientific evidence of the GOM is evaluated through its ability to improve the recognition accuracy of gestural time series that are modeled using continuous Hidden Markov Models (HMMs). Moreover,

the system is dynamically simulated through the solution of its equations. Its forecasting ability is evaluated by comparing the similarity between the real and the simulated motion data using two real observations for initializing the models as well as by measuring the Theil's inequality coefficient and its decompositions.

The performance of the algorithms that implement the GOM, the recognition of gestures and the forecasting of the motion trajectories is evaluated by recording four industrial real-life datasets from a European house-holding manufacturer, a glassblowing workshop, an AGV manufacturer and a scenario in automotive industry. More precisely, the first dataset contains motion data with gestures and actions from a TV assembly line, the second from the creation of glass water-carafes, the third gestural commands to mobile robots, and the fourth from a scenario of Human-robot collaboration in the automotive industry. The motion data used in our experiments are 2D positions that are exported from computer vision and the application of a deep-learning based pose estimation using the OpenPose framework [37].

Section 2 presents a State of the Art on human motion modeling, representation, and recognition. In Section 3, the whole methodology analysis, modeling and recognition is presented, while in Section 4 the different approaches in the evaluation of the ability of the models to simulate a gesture and forecast its trajectories are analysed. In the same section, the accuracy of the proposed method is also presented. Finally, in sections 5 and 6, a discussion and the future work and perspectives of the proposed methodology are described.

2 State of the Art

Movement can be defined as the change of someone's position, while gesture is a form of non-verbal communication used for controlling or interacting with a machine. Professional gestures define the routine of workers in industry. To reach the point of movements' interpretation, thus, to gesture recognition, it is essential to understand the existing relationships among human body parts.

2.1 Movement modeling and representation

Each body articulation is strongly affected by the movement of others. Observing a person running, brings evidence on the existing interdependencies between different parts of human body that need to move cooperatively for a movement to be achieved. Duprey et al.[1] attempted to study those relationships by exploring the upper body anatomy models available and describe their applicability using multi-body kinematic optimization, mostly for clinical and ergonomic uses. Biomechanics has also actively contributed to the study of human movement modeling by using Newtonian methods and approaches, especially in sports and physical rehabilitation [35]. Representing human movement with mathematical, physical and statistical models permits the estimation and forecasting of movement's evolution. This is also the goal of the State-Space, a statistical method that allows to forecast time series based on methods like Kalman filtering. A State-Space (SS) representation is a mathematical model of a physical system as a set of input, output and state variables related by first-order differential equations. Kalman filtering is a method that estimates and determines values for the parameters of a model. To represent human movement, Zalmai et al.[2], used linear SS models and provided an algorithm based on local likelihood for detecting and inferring gesture causing magnetic field variations. Lech and Kostek [3] used Kalman filtering to achieve hand tracking and presented a system based on camera and multimedia projector enabling a user to control computer applications by dynamic hand gestures. Finally, Dimitropoulos et al.[60] presented a methodology for the modeling and classification of multidimensional time series by exploiting the correlation between the different

channels of data and the geometric properties of the space in which the parameters of the descriptor lie by using a Linear Dynamical System (LDS). Here, multidimensional evolving data were considered as a cloud of points (instead of a single point) on the Grassmann manifold and we create a codebook in order to represent each multidimensional signal as a histogram of Grassmannian points, which is not always the case for professional gestures.

2.2 Machine learning for gesture recognition

Movement modeling and representation methods lead to gesture estimation but don't allow the modeling of precise movement patterns and consequently their recognition, as well as taking into consideration qualitative aspects of human movement such as expressivity. These limitations can be overcome with the use of machine learning methods. An important number of studies have been done in the past years in the field of gesture and movement recognition with the use of machine learning. Two different machine learning approaches have been mainly adopted to recognize various types of gestures: the model-based and the template-based methods.

2.2.1 Template-based Machine Learning

Template-based machine learning has been widely used for gesture recognition in the context of continuous real-time human-machine interaction. Dynamic Time Warping (DTW) is an example of methods that have been used to reach high gesture recognition accuracy results. DTW makes it possible to find the optimal global alignment between two sequences. Bevilacqua et al. [16], [17], [18] successively developed a system based on DTW, the Gesture Follower, for both continuous gesture recognition and following, between the template or reference gesture, and the input or performed one. A single example allows the training of the system [19]. During the performance, a continuous estimation of parameters is calculated in real-time, providing information for the temporal position of the performed gesture. Time alignment occurs between the template and the performed gesture, as well as an estimation of the time progression within the template in real-time. Instead of Psarrou et al. [11] used the Conditional Density propagation algorithm to perform gesture recognition, and make sure that they won't get probabilities for only one model per time-stamp. The experiments resulted to a relatively good accuracy for the time period conducted.

2.2.2 Model-based Machine Learning

One of the most popular methods of model-based machine learning that has been used to model and recognize movement patterns are Hidden Markov Models (HMMs). Pedersoli et al. adopted this method [4] to recognize in real-time static hand-poses and dynamic hand-gestures of American Sign Language. Sideridis et al. [6] created a gesture recognition system for everyday gestures recorded with Inertia Measurement Units, based on Fast Nearest Neighbors and Support Vector Machine methods while Yang and Sarkar [7] chose to use an extension of HMMs. Vaitkevičius et al. [10] used also HMMs with the same purpose, gesture recognition, for the creation of virtual reality installations, as well as Williamson and Murray-Smith [23] that used a combination of HMMs with a dynamic programming recognition algorithm, along with the granular synthesis method for gesture recognition with audio feedback. In a more industrial context, Yang, Park and Lee [25] used gesture spotting with HMMs to achieve efficient Human-Robot collaboration where real-time gesture recognition was performed with extended HMM methods like Hierarchical HMMs [8]. HMMs seem to be a solid approach allowing achieving satisfying results of gesture modeling and recognition and are suitable for real-time applications.

The aforementioned methodologies and research approaches permit the identification of what/which gesture is performed by giving a probability, but not how expressively the gesture has been performed. Caramiaux et al. [15] extended the research, by implementing a Sequential Monte Carlo technique to deal with expressivity. The recognition system, named Gesture Variation Follower, is being adapted to gesture expressive variations in real-time. Specifically, in the learning phase only one example per gesture is required. Then, in the testing (recognition) phase, time alignment is computed continuously and expressive variations (such as speed, size) are estimated between the template and the performed gesture [15][24].

The model-based and template-based methods present an interesting complementarity and their combination, in most of cases, give the possibility to achieve satisfying gesture recognition accuracy. However, when the probability given per class presents a high level of uncertainty, these methods need to be completed with an extra layer of control that will permit to take a final, more robust, decision about the probability of an observation to belong to each class. One of the goals of this work is to focus on the use of the State-Space method for human movement representation and modeling, and use this representation as the extra control layer to improve gesture recognition results.

2.2.3 Deep architectures for action recognition

Deep Learning (DL) is another approach with an increasing scientific evidence in action and gesture classification. Mathe et al. [27] presented results on hand gesture recognition with the use of a Convolutional Neural Network (CNN), which is trained on Discrete Fourier Transform images that were resulted from raw sensor readings. In [28], the authors proposed an approach for the recognition of hand gestures from the American Sign Language using CNNs and auto-encoders. 3DCNNs are used in [29] to detect hand movements of drivers and in [30] continuously recognize gesture classes from the Continuous Gesture Dataset (ConGD), which is the larger user-independent dataset. A two-stage approach is presented in [31], which combines feature learning from RGB-D using CNNs with Principal Component Analysis (PCA) for selecting the final features. Devineau, Xi, Moutarde and Yang [32] used a CNN model and tested its performance on classifying sequential humans' tasks using hand-skeletal data as input. Shahroudy et al. [57] wanting to improve their action recognition results and decrease the dependency in factors like lightning, background and color clothing, used a Recurrent Neural Network to model the long-term correlation of the features for each body part. For the same reason, Yan et al. [58] proposed a model of dynamic skeletons called Spatial-Temporal Graph Convolutional Network (ST-GCN). This Neural Network (NN) learns automatically the spatial and temporal patterns from the given data, minimizing the computational cost and increasing the generalization capability. In other cases, action recognition was achieved using either 3DCNNs [54] or two stream networks [59]. CNNs are the NNs used in all cases above, as they are the main method used for image pattern recognition.

The particularity of DL methods is that they require for a big amount of data in order to be trained. In some applications, having access to an important volume of data, might not be possible for various reasons. One application with extremely limited amount of data is the recognition of situated professional actions and gestures performed in an industrial context, such as in manufacturing, assembling lines, craftsmanship, etc. Deep NNs are powerful methods for pattern recognition with great accuracy results, but they present some limitations for real-time applications, which are linked to the computational power that is required both for training and testing purposes. In this paper, given the fact that the available examples per gesture class are also limited, it is assumed and proved that stochastic model-based machine learning can give better results than DL.

The aim of this paper is to get advantage of existing knowledge in machine learning, and more precisely in the stochastic modeling for the recognition of gestures and the forecasting of their motion trajectories, and compare their performance with a recent DL-based end-to-end architecture.

2.2.4 Our previous work

Manitsaris et al. [52] previously defined an operational model explaining how the body parts are related to each other, which was used for the extraction of confidence bounds over the time series of motion data. In [52], as well as in this from Volioti et al. [22], the operational model has been tested on Euler angles. In this work, the operational model is expanded to the full body and is tested with various datasets having different characteristics, e.g. more classes, more users and various real-life situations.

3 Methodology

3.1 Overview

The motion capturing of the operators in their workplace is a major task. A number of professional gestural vocabularies is created, to build the methodology and evaluate its scientific evidence. Although the proposed methodology (Figure 1) is compatible with various types of motion data, we opted for RGB sensors and, in most of cases, with 2D positions to avoid any interference between the operator and his/her tools or materials. Thus, RGB images are recorded for every gesture of the vocabulary, segmented into gesture classes, annotated, and then introduced to an external framework for estimating the poses and extracting the skeleton of the operators.

As shown in Figure 2, the GOM is based on a number of assumptions that describe the way the different entities of the human body cooperate to efficiently perform the gesture. The assumptions of the model refer to various relationships between the entities, which are: the *intra-joint association*, the *inter-limb synergies*, the *intra-limb mediation* and the *transitioning* over time. Following the theory of the SS modeling, the GOM is translated into a *simultaneous equations system* that is composed by two first-order differential equations for each component (e.g. dimension X, Y for 2D or X, Y, Z for 3D) of each body entity.

During the training phase, the motion data of the training dataset are used to compute the coefficients of the equations system using the MLE method but also to execute a supervised learning of the continuous HMMs. Moreover, the motion data are used to solve the simultaneous equations system and simulate the whole gesture, thus generating values for the state variables. Once the solution of the system is completed for all the gestures of the vocabulary, the forecasting ability of every model is evaluated using the Theil's coefficients as well as their performance in comparison with the motion data of other gestures.

During the testing phase, the HMMs output their likelihoods, which are multiplied by a confidence coefficient when their maximum likelihood is under a threshold. Finally, a motion trajectory can be dynamically or statically forecasted at any time by giving as input at least two time-stamps values from the real motion data.

3.2 Industrial datasets and gesture vocabularies

The performance of the algorithms is evaluated by recording four industrial real-life datasets from a house-holding manufacturer, a glassblowing workshop, an AGV manufacturer and an automotive industry. For each dataset, a gesture vocabulary has been defined in order to segment the whole procedure into small human motion units.

As shown in Table 2, the first gesture vocabulary (GV_1), includes 4 gestures where the operator takes the electronic card from one box, then takes a wire from another, connects them and places them on the TV chassis. The gestures are performed in a predefined working space, in front of the conveyor and with the boxes placed on the left and right side respectively. However, the operator has a certain degree of variation in the way of executing the tasks, since the gestures are ample involving the whole body. Moreover, in order to avoid self- and scene- occlusions, the camera is mounted on the top, which is not necessarily the optimal camera location for pose estimation algorithms, e.g. OpenPose. Currently in the factory, together with the operator who performs the actions of GV_1 there is also a second operator who will be progressively replaced by a collaborative robot.

The second dataset proposes gestural commands for controlling an AGV. As shown in Table 2, GV_2 contains five gestures involving mostly the arm and forearm. $G_{2,1}$ initiates the communication with the AGV, by shaking the palm, while $G_{2,2}$ and $G_{2,3}$ turn left and right the AGV by raising the respective arms. $G_{2,4}$ speeds up the AGV by raising three times the right hand, while $G_{2,5}$ speeds down the AGV by rolling the right hand away from the hips with a distance of around 20/30cm. All gestures of GV_2 start and end with the i-pose.

The third gesture vocabulary GV_3 contains four gestures performed by a glassblower when creating a water carafe (Table 2). The craftsman executes the gestures in a very limited space that is defined by a specific metallic construction. The craftsman puts the pipe on the metallic structure and to perform various manipulations of the glass by using tools, such as pliers, etc. The three out of four gestures are performed while the craftsman is sitting. More precisely, he starts by shaping the neck of the carafe with the use of pliers ($G_{3,1}$), then he tightens the neck to define the transition between the neck and the curved vessel ($G_{3,2}$), he holds in his right hand a specific paper and shapes the curves of the blown part ($G_{3,3}$) and finalizes the object and fixes the details by using a metallic stick ($G_{3,4}$). In general, the right hand is manipulating the tools while the left is holding and controlling the pipe. In parallel with $G_{3,2}$ and $G_{3,3}$, an assistant is helping and blowing promptly the pipe to permit the creation of the blown curved part.

The last dataset (GV_4) (Table 2) used in this paper, is related to a real-life Human-Robot Collaboration scenario that has been recorded in the automotive assembly lines of PSA Peugeot Citroën (PSA Group). A dual-arm robot and the worker are facing each other in order to cooperate for assembling motor hoses. More precisely, for the assembling of the motor hoses, the robot gives to the worker one part from the right and one part one from the left claw, the worker takes two hose parts from the robot, joins them, screws them, and finally places the mounted motor hose in a box. In order for the robot to achieve the appropriate level of perception and move accordingly, it needs to make two specific actions “to take a piece in the right claw” and “to take a piece in the left claw”. Then, the worker can screw after the first gesture “to assemble”, or can choose to screw later during the last assembly sub-task. At the end of the assembly task, the worker puts the assembled piece in a box, which means that a cycle has just ended. Therefore, it is important for the robot to recognize the actions “to assemble” and “to screw” of the worker, so as to give at the correct moment the next motor piece with its arm. Twelve operators have been recorded in GV_4 .

The four datasets and vocabularies contain professional gestures performed in different industries and contexts. Important differences may be observed between them though. For example, GV_1 , GV_3 and GV_4 involve the manipulation of tools from the operator. Therefore, the distribution of variances alternates between high, e.g. when moving for grabbing the tool, and low values, e.g. when tools or objects are put on a specific position. In GV_1 , despite the fact that the gestures are performed in a predefined space, the operator has a certain degree of variation between different repetitions of the same gesture. Human

factors such as the level of experience, fatigue or even stress, influence the way these gestures are performed without necessarily having a direct impact on the final result, which is to place the card on the TV. However, this is not the case of the GV_3 where a high level of technicity and dexterity is required. In GV_3 , only a low spatial and temporal variation can be accepted. The glass blower performs the gestures with a high repeatability from one repetition to another and successfully reproduce the object with exactly the same specifications, e.g. size, diameter etc. The gestural commands of GV_2 are simpler and ampler. A bigger freedom and variation are thus expected in the way they are performed. In GV_4 , the operator is performing actions with a high repeatability. Since the dual-arm robot Motoman SDA20 has been used, the operator, depending on if he/she is left or right-handed, has various possibilities for grabbing the parts from the robot and the tools.

In GV_1 , while all the gestures are performed by a single user, the different position of the operator in space in each gesture makes it an interesting dataset to work on. Also, this dataset appears to have a lot of noise, and it was an opportunity to examine the reaction of the pose estimation framework to noisy data. The second dataset (GV_2) has multiple users, giving the opportunity to examine how gesture recognition works with a high variation among the performance of the same gestures. In the third gestural vocabulary (GV_3), all gestures have been performed by an expert artist. They are fine movements where hands are cooperating in a synchronous way. Consequently, investigating body parts dependencies in this GV becomes extremely interesting. The fourth gestural vocabulary (GV_4), has a robot involved in the industrial routine.

From an *intra-class* variance point of view, the Root-Mean-Square-Error is used to evaluate the datasets. The root-mean-square-error (RMSE) allows the measurement of the difference between two times-series and it is defined as shown in equation 1.

$$\text{RMSE} = \sqrt{(o_1 - o_2)^2} \quad (1)$$

where o_1 is one of the iterations of a specific gesture within a gestural vocabulary and o_2 is another iteration of the same gesture, among which the variance is to be examined. A high variance between the iterations of each gesture of GV_2 is noticed, which is the expected result, since this dataset consists of gestures performed by six different users (Table 3). The RMSE for GV_3 appears to have low *intra-class* variation, as expected, since the gestures are performed by an expert, who is able to repeat them in a very precise way.

3.3 Pose estimation and features extraction

After the motion capturing and recording of the data, each image sequence of the three first datasets, is imported to the OpenPose framework, which detects body keypoints on the RGB image and extracts a skeletal model together with the 2D positions of each body joint [61] (Figure 2). These joints are not necessarily physical joints. They are keypoints on the RGB image which, in most cases, correspond to physical joint centers. OpenPose uses the neck as the root keypoint to compute all the other body keypoints (or joints). Thus, the motion data are normalized by using the neck as the reference joint. In addition to this, the coordinates of each joint are derived by the width and height of the camera. With regard to GV_4 , 3D hand positions are extracted from top-mounted depth imaging by detecting keypoints on the depth map. The keypoints are localized by the computing the geodesic distances between the closest body part to the camera (head) and the farthest visible body part (hands), as it is presented in our previous work [34]. Any vision-based pose estimation framework may output 2D positions of a low precision, depending on the location of the camera, such as OpenPose for a top-mounted view.

However, these errors may not strongly affect the recognition accuracy of our hybrid approach. This is also proved by the fact that our approach outperforms the End-to-End 3DCNNs that doesn't use any skeletisation of the human body to recognize the human actions.


The extracted features for each joint, were the X and Y positions, as they are provided by OpenPose. More specifically, for GV_1 the 2D positions of the two wrists have been used, while for GV_2 and GV_3 , the 2D positions of the head, neck and shoulder, elbows, wrists and hands, as they were proven to give optimal recognition results. With regard to GV_4 , 3D hand positions are used.

3.4 Gesture Operational Model


When a skilled individual performs a professional situated gesture, the whole body is involved combining thus theoretical knowledge with practical motor skills. Effective and accompanying body movements are harmonically coordinated to execute a given action. The expertise in the execution of professional gestures is characterized by precision and repeatability while the body is continuously shifting from one phase to another, e.g. from specific postures (small tolerance for spatial variance) to ample movements (high tolerance for spatial variance). For each phase of the movement, each body entity, e.g. articulation or segment, moves in a multidimensional space over time. When considering the 2D motion descriptors of the movement, two mutually depended variables represent the entity, e.g. X and Y positions. Each of these variables is associated with the other, creating thus a bidirectional relationship between them. Furthermore, they also depend on their history while some entities might 'work together' to execute an effective gesture, e.g. when an operator assembly two parts. However, a unidirectional dependency might be observed when one entity influences the other entity and not vice versa as well as a bidirectional dependency when both entities influence one each other, e.g. when a potter shapes the clay with both hands.

The above observations on situated body movements can be translated into a functional model, that we define here as the Gesture Operational Model (GOM), which describes how the body skeletal entities of a skilled individual, are organized to deliver a specific result (Figure 2). It is assumed that each of the assumptions of '*intra-joint association*', '*transitioning*', '*intra-limb synergies*' and '*intra-limb mediation*', contribute at a certain level to the production of the gesture. As far as the *intra-limb mediation* is concerned, it can be decomposed into the '*inter-joint serial mediation*' and the '*inter-joint non-serial mediation*'. The proposed model works perfectly for all three dimensions (X, Y and Z), but for simplicity reasons, it will be presented only for two dimensions, the X and Y. In addition to this, in this work only positions are used, but the model is designed to be able to receive joint angles as input as well.

H1: Intra-joint association

It is hypothesized that the motion of each body part (*Entity*) (e.g. right hand) is decomposed in a motion on X-axis and Y-axis, thus described by two mutually depended variables. It is assumed that there is a bidirectional relationship between the two variables, defined here as *intra-joint assumption* and indicated by .

H2: Transitioning

It is also assumed that each variable depends on its own history, also called inertia effect. This means that the current value of each variable depends on the values of previous times, also called lag or dynamic effect, which is defined here as *transitioning* and indicated by .

H3: Inter-limb synergies

It is assumed that some body entities, work together to achieve certain motion trajectories, e.g. hands when assembling two parts, defined here as *inter-limb synergies*.

H4: Intra-limb mediation

Inter-joint serial mediation: It is assumed that a body entity may depend on its neighboring entities to which it is directly connected to, e.g. a glassblower, while using the pipe, moves his/her wrists along with his/her shoulders and elbows. In case this assumption is statistically significant there is an *inter-joint serial mediation*.

Inter-joint non-serial mediation: It is assumed that each body entity depends on non-neighboring entities of the same limb, e.g. the movement of the wrist may depend on the movement of the elbow and shoulder. Thus, it is highly likely that both direct and indirect dependencies simultaneously occur in the same gesture. *Entities* are named after the first letters of the respective body joint. More specifically, LSH and RSH represent the left and right shoulder respectively. Accordingly, LELBOW and RELBOW represent the left and right elbow, LWRIST and RWRIST, the left and right wrist, LHAND and RHAND the left and right hand. HEAD, NECK and HIPS represent, as their names indicate, the head, the neck and the hips.

So, an example of the representation of those assumptions for the X-axis would be a below:

$$Entity_{1,x}(t) = Entity_{1,y}(t - 1) + Entity_{1,x}(t - 1) + Entity_{1,x}(t - 2) + Entity_{2,x}(t - 1) \quad (2)$$

3.5 Simultaneous Equations System

The simultaneous equations system concatenates the dynamics of an N^{th} order system, the GOM, into N first-order differential equations. The number of equations is equal to the number of associated dimensions to a given entity multiplied by the number of body entities. Therefore, the steps to follow are the *estimation* of the model, with the aim of verifying its structure, as well as the *simulation* of the model to verify its forecasting ability.

3.5.1 State-Space representation

The definition of the equations of the system follows the theory of the SS modeling, which gives the possibility for the coefficients to dynamically change over time. A SS model for n -dimensional time series $y(t)$, consists of a *measurement or observation equation* relating the observed data to an m -dimensional state vector $s(t)$ and a Markovian *state or transition equation* that describes the evolution of the state vector over time. The *state equation* depicts the dependence between the system's past and future and must 'canalize' through the state vector. The *measurement or observation equation* is the 'lens' (signal) through which the hidden state is observed and it shows the relationship between the system's state, input and output variables. Representing a dynamic system in a SS form, allows the state variables to be incorporated into and estimated along with the observable model.

Therefore, given an input $u(t)$ and a state $s_S(t)$, a SS gives the hidden states that result to an observable output (signal). A general SS representation is as follows:

$$\frac{ds_S}{dt} = As_S(t - 1) + w(t) \quad (3)$$

$$y = C \frac{ds_s}{dt} + Du \quad (4)$$

where (3) is the *state* equation, which is a first-order Markov process (4) is the *measurement* equation, s_s is the vector of all the state variables, $\frac{ds_s}{dt}$ is the time derivative of the state vector, u is the input vector, y is the output vector, A is the transition matrix that defines the weight of the precedent space, C is the output matrix and D is the feed-through matrix that describes the direct coupling between u and y , and t indicates time.

When capturing the gestures with motion sensors, Gaussian disturbances are also added in both the state and the output equation. After performing the experiments presented in this work, it was observed that Gaussian disturbances didn't change at all the final estimation result, so they were considered to be negligible.

The SS representation of the positions on the X-axis for a body *Entity_{i,j}* -where i represents the body part modeled in a SS form and j the dimension of each Entity- according to the GOM is structured as follows:

$$\frac{ds_s}{dt} = A * s_s(t-1) = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} Entity_{1,X}(t-1) \\ -Entity_{1,X}(t-2) \end{bmatrix} = \begin{bmatrix} \alpha_1 Entity_{1,X}(t-1) \\ -\alpha_2 Entity_{1,X}(t-2) \end{bmatrix} \quad (5)$$

$$\begin{aligned} \stackrel{(5)}{\Rightarrow} Entity_{1,X}(t) &= [1 \ 0] \frac{ds_s}{dt} + \alpha_3 Entity_{1,Y}(t-1) + \alpha_4 Entity_{2,X}(t-1) = \\ &= \alpha_1 Entity_{1,X}(t-1) - \alpha_2 Entity_{1,X}(t-2) + \alpha_3 Entity_{1,Y}(t-1) + \alpha_4 Entity_{2,X}(t-1) \end{aligned} \quad (6)$$

Where α_i , the coefficients that need to be estimated. In equation 6, $Entity_X(t-2)$ is subtracted by $Entity_X(t-1)$, indicating difference between successive levels of dimensions, e.g. positions on Y-axis (transitioning assumption). Equations 5 and 6 occur by equations 3 and 4 respectively. More specifically, equation 6 consists of the exogenous variables to which the endogenous ones, coming up from the state equation (equation 5), are added.

Equation 6 has the form of a first-order Autoregressive (AR) model. An AR model predicts future behavior based on past behavior. The order of the autoregressive model is adapted in each case according to the data characteristics and the experiments. During the performance of the experiments, the use of an autoregressive model of second order led to better estimation results. As such, in the transitioning assumption, the position values of the two previous time periods (frames) of a given axis are considered.

For the modeling of the full human body, the simultaneous equations system is based on the equations 5 and 6, which consist of 2 sets of equations for each used entity, one for each dimension X, and Y. Thus, for a full body GOM, we obtain 32 equations describing 32 endogenous variables with 64 state variables that contain 2 exogenous variables for $t-1$ and $t-2$.

As an example, the SS representation for the right wrist is given:

$$RWRIST_X(t) = \alpha_1 RWRIST_X(t-1) - \alpha_2 RWRIST_X(t-2) + \alpha_3 RWRIST_Y(t-1) + \alpha_4 LWRIST_X(t-1) \quad (7)$$

In equation 7, $RWRIST_X(t-1)$ and $RWRIST_X(t-2)$ are the endogenous variables, while $RWRIST_Y(t-1)$, and $LWRIST_X(t-1)$ are the exogenous ones.

3.6 Computing the coefficients of the equations

The coefficients of the simultaneous equations system are computed using the MLE method via Kalman filtering [45]. Let consider a gesture $G_{j \in \mathbb{N}}$ of a gesture vocabulary $GV_{i \in [1,3]}$ and an observation $\mathcal{O}_{0:k} = \{o_1 \dots o_k\}_{k \in \mathbb{N}}$, where o_k is one observation vector and k the total number of observations. Thus, the probability \mathcal{P}_s to observe o_t at time $t \in [0, k]$ will be as follows:

$$\mathcal{P}_s(\mathcal{O}_{0:k}) = \prod_{t=0}^k \mathcal{P}(o_t | \mathcal{O}_{0:t-1}) \quad (8)$$

where k represents the observed data, $\mathcal{P}(o_t | \mathcal{O}_{0:t-1})$ is the probability of o_t given all the observations before time t .

Also, the probability of time series given a set of parameters Ψ , is

$$\mathcal{P}(\mathcal{O}_{0:t-1} | \Psi) = \prod_{t=1}^k \exp \left\{ -\frac{(o_t - \tilde{o}_t^{t-1})^2}{2F_t^{t-1}} \right\} (2\pi |F_t^{t-1}|)^{-\frac{1}{2}} d\theta \quad (9)$$

with variance F_t^{t-1} and mean \tilde{o}_t^{t-1} . So, the log-likelihood of ψ given data $\mathcal{O}_{0:t-1}$ is

$$\log L(\Psi | \mathcal{O}_{0:t-1}) = -\frac{k}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^k \log |F_t^{t-1}| - \frac{1}{2} \sum_{t=1}^k \frac{(o_t - \tilde{o}_t^{t-1})^2}{F_t^{t-1}} \quad (10)$$

For the computation of this log-likelihood, the estimation, the variance and mean that appear in equation 4, need to be estimated optimally. Kalman filtering gives the optimal estimates of the mean and covariance for the calculation of the maximum likelihood of ψ . Kalman filtering consists of two main recursive steps, prediction and update. In the first step, there is an estimation of the mean and covariance, along with the predicted error covariance. In the update step, the optimal Kalman gain is computed, so the estimation of mean and covariance from the prediction step is updated according to it. These two steps appear recursively, until the optimal \tilde{o}_t^{t-1} and F_t^{t-1} that fit the observed data, are computed. This derives the computation of the coefficients of the SS equations and the forecasting of a new time series given those observed data.

3.7 Learning with Hidden Markov Models

HMMs follow the principles of Markov chains that describe stochastic processes. They are commonly used to model and recognize human gestures. They are structured using two different types of probabilities, the transition probability from one state to another and the probability for a state to generate specific observations on the signal [46]. In our case, each professional gesture is associated to an HMM, while the intermediate phases of the gesture constitute internal states of the HMM. According to our datasets, these gestures define the gesture vocabulary $GV_{i \in [1,3]} = \{G_j\}_{j \in \mathbb{N}}$.

Let \mathcal{S}_h be a finite space of states, corresponding to all the intermediate phases of a professional gesture. The transition probability between the states $\mathcal{Q}(s_h, s'_h)$ where $s_h, s'_h \in \mathcal{S}_h$ are given in the transition matrix $\mathcal{Q} = [\mathcal{Q}(s_h, s'_h)]$. A hidden sequence of states where $\mathcal{S}_{h0:k} = \{s_{h1} \dots s_{hk}\}_{k \in \mathbb{N}}$ where $s_{hk} \in \mathcal{S}_h$ is also considered. A given sequence of hidden states $\mathcal{S}_{h0:k}$ is supposed to generate a sequence of observation vectors $\mathcal{O}_{0:k}$. We assume that the vectors \mathcal{O}_k depends only on the state s_{hk} . From now on, the likelihood that the observation \mathcal{O} is the result of the state s_h , will be defined as $\mathcal{P}_h(\mathcal{O}|s_h)$. It is important to outline that in our modeling structure; each internal state of the model depends only on its previous state (first-order Markov property). Consequently, the set of the models for all gestures for every gesture vocabulary is $GV_{j \in [1,3]} = \{\text{HMM}_i\}_{i \in \mathbb{N}}$, where $\text{HMM}_i = (\mathcal{Q}_i, \mathcal{Q}_i, \mathcal{P}_{hi})_{i \in \mathbb{N}}$ are the parameters of the model and \mathcal{Q}_i is the initial state probability. Thus, the recognition becomes an issue of solving three specific problems: *evaluation*, *recognition* and *learning* [47]. Each one of those problems was solved with the use of the algorithms, Viterbi [43], Baum's "forward" [38] and Baum-Welch respectively [44].

3.8 Gesture recognition

In the recognition phase, the main goal is to recall with a high precision the hidden sequence of internal states $\mathcal{S}_{h0:k}$ that correspond to the sequences of the observation vectors. Thus, let consider the observation of motion data $\mathcal{O}_{0:k}$, which need to be recognized. Every HMM_λ with $\lambda \in [1, j]$ of a given GV_i with $i \in [1, 3]$ generate the likelihood $\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|\text{HMM}_\lambda)$. If there is at least one HMM_ξ with $\xi \in [1, j]$ that generates $\mathcal{P}_{h\xi} \geq 0.55$ then it is considered that $\mathcal{O}_{0:k}$ is generated by $G_{i,\xi}$. Otherwise, the following quantity is computed for every SS_λ of GV_i (confidence control):

$$SS_\lambda^{\text{score}} = \frac{1}{1 + d(\mathcal{O}_{0:k}, \mathcal{O}_{0:k,\lambda}^s)} \quad (11)$$

where d is the minimum distance between the simulated values $\mathcal{O}_{0:k,\lambda}^s$ from the model SS_λ and the original observations $\mathcal{O}_{0:k}$.

Then, for every SS_λ of GV_i the likelihood $\mathcal{P}'_{h\lambda}(\mathcal{O}_{0:k}|\text{HMM}_\lambda^{\text{SS}})$ is computed as follows:

$$\mathcal{P}'_{h\lambda}(\mathcal{O}_{0:k}|\text{HMM}_\lambda^{\text{SS}}) = \mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|\text{HMM}_\lambda) \cdot SS_\lambda^{\text{score}} \quad (12)$$

Therefore, the final formula providing the way the algorithm recognizes the observation of motion data $\mathcal{O}_{0:k}$

$$\mathcal{R}_{GV_i}(\mathcal{O}_{0:k}) = \begin{cases} \max_1^j(\mathcal{P}_{hi}(\mathcal{O}_{0:k}|\text{HMM}_i)), & \max(\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|\text{HMM}_\lambda)) \geq 0.55 \\ \max_1^j(\mathcal{P}'_{h\lambda}(\mathcal{O}_{0:k}|\text{HMM}_\lambda^{\text{SS}})), & \max(\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|\text{HMM}_\lambda)) < 0.55 \end{cases} \quad (13)$$

4 Evaluation

The evaluation of the accuracy and performance of the method follows an 'all-shots' approach for the training of the Hidden Markov Models and an 'one-shot' approach for estimating the coefficients of the State-Space models.

In order to select which gestural iteration to use for computing the coefficients of the State-Space models, the "leave one out method" is used. It is a resampling technique which is also useful for

variance and bias estimation (and avoidance), especially when the data are limited. It consists in systematically leaving out one observation from a dataset, calculating the estimator and then finding the average of these calculations. In our case, the estimator was the likelihood of the HMM when trained with one iteration of a gesture and tested with all the other iterations. The iteration giving the maximum likelihood is selected for computing the coefficients of the State-Space models.

4.1 Statistical significance and simulation of the models

In order to evaluate the significance of the assumptions concerning the body parts dependencies that are defined within the GOM, a statistical significance analysis is done. The statistical significance p-value indicates whether the assumptions are verified or not. The level of statistical significance is often expressed by using the p-value, which takes values between zero and one. Generally, the smaller the p-value, the stronger the evidence that the null hypothesis should be rejected. In this work the 0.05 p-value was used as the threshold for the statistical significance tests. If the p-value of the estimated coefficient is smaller than 0.05, then the specific coefficient is statistically significant and need to be included in the SS representation of the model.

In the case of the professional gestures, investigating the significance level of the coefficients of each variable within the GOM, explains how important is each joint for each gesture in the gestural vocabulary. Examples of some of the gestures from GV_2 and GV_3 are given, to observe cases where some of the coefficients affect strongly the results and need to remain dynamic, while others not, and can remain constant. In the GOM below, the equation of $G_{2,1}$ for $RWRIST_X$ is as follows, starting from equation 2.

$$\begin{aligned} RWRIST_X(t) &= a_{12}RWRIST_Y(t-1) + a_{13}RWRIST_X(t-1) - a_{14}RWRIST_X(t-2) + a_{15}LWRIST_X(t-1) = \\ &= \overbrace{-0.0629}^{0.266} RWRIST_Y(t-1) + \overbrace{1.3438}^{0.00} RWRIST_X(t-1) - \left(\overbrace{-0.3648}^{0.00} \right) RWRIST_X(t-2) + \left(\overbrace{-0.6625}^{0.449} \right) LWRIST_X(t-1) \end{aligned} \quad (14)$$

Having performed the statistical significance analysis of the model in equation 14, we get the estimation of the coefficients. Where equation 14 is the general equation for X-axis of the right wrist, along the p- values that indicate the level of significance of each part of the equation. The p-values show that in the case of the $G_{2,1}$, the past values on the same axis appear to be significant, while the respective p-values of the left wrist or the Y-axis of the right wrist, are not statistically significant. This result was expected, as this gesture is a ‘hello waving movement’, where the right wrist is moving across the X axe and the left wrist remains still through the performance of the gesture, leading to the result that there is no intra-limb mediation in this specific gesture.

Following, there is one more example of the same GV , from gesture $G_{2,3}$, for X-axis (equation 15) and Y-axis (equation 16). The numbers above the estimated coefficients correspond to their respective p-values.

$$\begin{aligned} RWRIST_X(t) &= a_{12}RWRIST_Y(t-1) + a_{13}RWRIST_X(t-1) - a_{14}RWRIST_X(t-2) + a_{15}LWRIST_X(t-1) = \\ &= \overbrace{-0.2871}^{0.00} RWRIST_Y(t-1) + \overbrace{0.6392}^{0.00} RWRIST_X(t-1) - \overbrace{0.0273}^{0.86} RWRIST_X(t-2) + \overbrace{0.0516}^{0.00} LWRIST_X(t-1) \end{aligned} \quad (15)$$

$$\begin{aligned}
 \text{RWRIST}_Y(t) &= a_{12}\text{RWRIST}_X(t-1) + a_{13}\text{RWRIST}_Y(t-1) - a_{14}\text{RWRIST}_Y(t-2) + a_{15}\text{LWRIST}_Y(t-1) = \\
 &= \overbrace{-3.9907}^{0.00} \text{RWRIST}_X(t-1) + \overbrace{0.5003}^{0.00} \text{RWRIST}_Y(t-1) - \overbrace{(-0.0818)}^{0.616} \text{RWRIST}_Y(t-2) + \overbrace{(-0.0927)}^{0.00} \text{LWRIST}_Y(t-1)
 \end{aligned} \tag{16}$$

In this gesture, the operator moves his right wrist towards his right side both on the X and Y axes, indicating to the AGV to turn right. So, according to the results, all coefficients appear to be statistically significant, apart from the 2 previous time-periods value of the X-axis of the right wrist. The same results occur for the Y-axis of the same wrist.

To verify the results, a significance level test is presented for $G_{3,2}$ of GV_3 . During the performance of this gesture, the glassblower is moving both wrists cooperatively, to tighten the base of the glass piece. The right wrist works more intensively to complete tightening the glass, while the left wrist complements the movement by slowly rolling the metal pipe.

$$\begin{aligned}
 \text{RWRIST}_X(t) &= a_{12}\text{RSH}_X(t-1) + a_{13}\text{RELBOW}_X(t-1) + a_{14}\text{RWRIST}_Y(t-1) + a_{15}\text{LWRIST}_X(t-1) + \\
 &\quad + a_{16}\text{RWRIST}_X(t-1) + a_{17}\text{RWRIST}_X(t-2) = \\
 &= \left(\overbrace{-0.0778}^{0.562} \right) \text{RSH}_X(t-1) + \overbrace{1.1126}^{0.00} \text{RELBOW}_X(t-1) + \left(\overbrace{-0.4757}^{0.00} \right) \text{RWRIST}_Y(t-1) + \overbrace{0.3423}^{0.00} \text{LWRIST}_X(t-1) + \\
 &\quad + \overbrace{0.4585}^{0.00} \text{RWRIST}_X(t-1) + \overbrace{0.4604}^{0.00} \text{RWRIST}_X(t-2)
 \end{aligned} \tag{17}$$

$$\begin{aligned}
 \text{RWRIST}_Y(t) &= a_{12}\text{RSH}_Y(t-1) + a_{13}\text{RELBOW}_Y(t-1) + a_{14}\text{RWRIST}_X(t-1) + a_{15}\text{LWRIST}_Y(t-1) + \\
 &\quad + a_{16}\text{RWRIST}_Y(t-1) + a_{17}\text{RWRIST}_Y(t-2) = \\
 &= \overbrace{0.290}^{0.117} \text{RSH}_Y(t-1) + \overbrace{0.3678}^{0.00} \text{RELBOW}_Y(t-1) + \left(\overbrace{-1.0912}^{0.00} \right) \text{RWRIST}_X(t-1) + \left(\overbrace{-0.1602}^{0.045} \right) \text{LWRIST}_Y(t-1) + \\
 &\quad + \overbrace{1.1298}^{0.00} \text{RWRIST}_Y(t-1) + \overbrace{(-0.1679)}^{0.00} \text{RWRIST}_Y(t-2)
 \end{aligned} \tag{18}$$

$$\begin{aligned}
 \text{LWRIST}_X(t) &= a_{12}\text{LSH}_X(t-1) + a_{13}\text{LELBOW}_X(t-1) + a_{14}\text{LWRIST}_Y(t-1) + a_{15}\text{RWRIST}_X(t-1) + \\
 &\quad + a_{16}\text{LWRIST}_X(t-1) + a_{17}\text{LWRIST}_X(t-2) = \\
 &= \overbrace{0.3668}^{0.00} \text{LSH}_X(t-1) + \overbrace{0.11180}^{0.007} \text{LELBOW}_X(t-1) + \overbrace{0.9589}^{0.00} \text{LWRIST}_Y(t-1) + \left(\overbrace{-0.0126}^{0.339} \right) \text{RWRIST}_X(t-1) + \\
 &\quad + \overbrace{1.1111}^{0.00} \text{LWRIST}_X(t-1) + \overbrace{(-0.1398)}^{0.052} \text{LWRIST}_X(t-2)
 \end{aligned} \tag{19}$$

$$\begin{aligned}
 \text{LWRIST}_Y(t) &= a_{12}\text{LSH}_Y(t-1) + a_{13}\text{LELBOW}_Y(t-1) + a_{14}\text{LWRIST}_X(t-1) + a_{15}\text{RWRIST}_Y(t-1) + \\
 &\quad + a_{16}\text{LWRIST}_Y(t-1) + a_{17}\text{LWRIST}_Y(t-2) = \\
 &= \overbrace{0.00}^{0.00} \text{LSH}_Y(t-1) + \overbrace{0.5433}^{0.00} \text{LELBOW}_Y(t-1) + \overbrace{0.1144}^{0.272} \text{LWRIST}_X(t-1) + \overbrace{0.0162}^{0.356} \text{RWRIST}_Y(t-1) + \\
 &\quad + \overbrace{1.0463}^{0.0} \text{LWRIST}_Y(t-1) + \overbrace{(-0.1095)}^{0.124} \text{LWRIST}_Y(t-2)
 \end{aligned} \tag{20}$$

In the equations presented above, all coefficients appear to be statistically significant, except from $\text{RSH}_X(t-1)$ in equation 17, $\text{LWRIST}_X(t-1)$ in equation 18, $\text{LWRIST}_X(t-1)$ and $\text{RWRIST}_X(t-2)$ in equation 19, $\text{RWRIST}_X(t-1)$, $\text{RWRIST}_Y(t-1)$, $\text{LWRIST}_X(t-1)$ in equation 20. As a result, the hands of the operator work mostly independent (there appears to be a dependency in the inter-limb synergies in equation 16), while all the other assumptions seem to be statistically significant for both X-axis and Y-axis of the right and left wrist.

The simulation of the models is based on the solution of their simultaneous equations system. Figures 3, 4 and 5 show examples of the graphical depiction of real observations of motion data together with their simulated values from the State-Space model of the right wrist. A general conclusion that can be exported by looking at the depictions is that the behavior of the models is very good because the two curves are really close in most cases.

4.2 Recognition performance of professional gestures and comparison with end-to-end deep learning architectures

For the evaluation of the performance and the proposed methodology, the metrics *precision*, *recall* and *f – score* were calculated. Those metrics are defined as shown below.

$$\text{precision} = \frac{\#(\text{true positives})}{\#(\text{true positives}) + \#(\text{false positives})} \tag{21}$$

$$\text{recall} = \frac{\#(\text{true positives})}{\#(\text{true positives}) + \#(\text{false negatives})} \tag{22}$$

Precision, *recall* and *f – score* are calculated for all the gestures that each gestural vocabulary consists of. For a gesture of class i , $\#(\text{true positives})$ represent the number of gestures of class i that were recognized correctly, $\#(\text{false positives})$ represent the number of gestures that didn't belong in class i and they were recognized from the algorithm as parts of class i . Finally, $\#(\text{false negatives})$ represents the number of gestures belonging to class i that were not recognized as part of it.

More precisely, *precision* represents the rate of gestures that really belong in class i , among those who are recognized as class i , while *recall* represents the rate of iterations of gestures of class i that have

been recognized as class i . A measure that combines both precision and recall is the f – score, which is given by equation (23).

$$f - score = 2 \frac{precision * recall}{precision + recall} \quad (23)$$

The performance of the algorithms was tested with the four different gestural vocabularies. As presented before the GV_1 contains 4 classes, from 44 to 48 repetitions for each. 4 hidden states were used for HMMs training. To simplify the evaluation task, a simplified GOM with only X and Y positions of two wrists are used for training and recognition. Table 4 presents the results when only the HMMs are used for recognition without any confidence control, and also the results with the confidence control provided by the simulation of the State-Space models.

It is possible to observe that HMMs provide a recall superior to 90% in 3 out of 4 gestures. The $G_{1,3}$ presents the lowest recall of 81.81 % and this can be due to the fact that this is the most complex gesture, where both hands interact more than in the other 3 gestures. The lowest precision is detected for the $HMM_{1,4}$. When the SS representation and confidence control is used, the *recall* for $G_{1,2}$ is slightly improved while in the case of the $G_{1,3}$ a significant improvement of 15,91% is achieved. Especially for $G_{1,3}$, the improvement can be justified by the fact that the operator is connecting the wire with a very small card outside the conveyor. Thus, the operator has the possibility to perform very small movements in different positions of his/her workplace. The *precision* of $HMM_{1,4}^{SS}$ has been also positively impacted by the SS augmenting the accuracy from 90,2% to 97,67. However a slight decline can also be seen in the case of $G_{1,4}$ recall. (Table 3)

The GV_2 contains 5 classes, 16 repetitions of each gesture and 1-11 hidden states were used for the machine learning gesture recognition engine according to the best states' number for each iteration. The joints selected for training with GV_2 were the wrist, elbow and shoulder joints for each hand, along with the neck. In Table 5 *precision* and *recall* using only HMM and the HMM^{SS} approach is presented respectively. For $G_{2,1}$, $G_{2,4}$ and $G_{2,5}$ ergodic topology was used, as iterations of the same gestural part appear during the performance of each gesture, while left to right topology was used for the rest of the gestures. *Precision* appears improved for every model, while *recall* is decreased for $G_{2,2}$ and $G_{2,5}$. (Table 4)

GV_3 consists of 4 different gestures with 35, 34, 21 and 27 repetitions respectively. 5-20 hidden states were used for training the gesture recognition algorithm, the number of which were again computed for every iteration in the resampling phase. The joints selected for training with GV_3 were again the wrist, elbow and shoulder joints for each hand, along with the neck. *Precision* appears improved in almost every observation and maximum likelihood. The *recall* in almost every gesture has remained stable except from $G_{3,3}$ where it was increased by +4%. (Table 5)

GV_4 consists of five different gestures. The clusters used in the k-means approach in combination with an HMM with 12 hidden states, were 25. The proposed methodology in this work performed better than the rest machine learning methods, with f – score results improved by +12%.

In Table 7, the comparison of mean f – scores for each GV and each approach is presented. The score of GV_1 and GV_2 was improved by approximatively 2% while the most important contribution is observed for the GV_3 . The HMM^{SS} allow to improve significantly (+7.5%) the recognition results of this last dataset.

A similar conclusion can be extracted from the same table, where the total accuracy for the GV_3 has reached 80.34% from 70.94%. The accuracy improvement of the two other datasets remain at the same level with the one of the mean $f - score$, around +2%.

In order to compare the results of the approach proposed in this paper with other classification techniques, a Deep Learning End-to-End 3D Convolutional Neural Network has been used to classify the gestures of the three first vocabularies described in the section 3.2. More precisely a 3DCNN has been initially trained on spatiotemporal features from a medium sized UCF-101 video dataset and the pretrained weights have been used to finetune the model on small sized datasets including images of operators performing customized gestures in industrial environments.

The architecture of the network was based on the one proposed in [54] with 4 convolution and 2 pooling layers, 1 fully-connected layer and a softmax loss layer to predict action labels. It has been trained from scratch on the UCF-101 video dataset, using batch size of 32 clips and Adam optimizer [55] for 100 epochs, with Keras deep learning framework[56]. The entire network was frozen and only 4 last layers were finetuned on customized datasets by backpropagation.

The comparison of recognition accuracy results between HMMs, HMM^{SS} and 3DCNN is shown in the tables 3, 4, 5, 6 and 7. As far as the GV_1 is concerned, the use of a 3DCNN improves the recognition of only one gesture ($G_{1,1}$) as shown in the table 3. However, in total the HMM^{SS} outperform the other two methods reaching a total accuracy of 96.19% (Table 7). In the second dataset (GV_2), 3DCNN doesn't achieve a satisfying recognition result for the $G_{2,5}$ (66%) in comparison to other methods that reach 100% (Table 4) and in total HMM^{SS} still performs the best as it is possible to observe in the Table 7. In the GV_3 , the HMM^{SS} performs again the best among the three methods, as shown also in Table 7, with a total accuracy almost +4% higher and an $f - score$ of +1.5% higher than the DL method.

4.3 Forecasting ability for motion trajectories

For the evaluation of the ability of the 4 State-Space models that are used to explain the assumptions of the two-entities GOM, a simulation using equation (2) for all three dimensions and for all used joints was performed (Table 8). It includes the computation of Theil's inequality coefficient (U) and its decomposition into the inequality of bias proportion U^B , variance proportion U^V and covariance proportion U^C . U^B examines the relationship between the means of the actual values and the forecasts, U^V considers the ability of the forecast to match the variation in the actual series and U^C captures the residual unsystematic element of the forecast errors [48]. Thus, $U^B + U^V + U^C = 1$. Theil inequality coefficient measures how close the simulated variables are to the real variables and it gets values from 0 to 1. The closer to zero the value of this factor is, the better the forecasting of the variable. Also, the forecasting ability of the model is better, when U^B and U^V are close to zero and U^C is close to one. The computed coefficients as shown in Table 8, result to a sufficient forecasting performance of the simulated model and the error results reinforce this conclusion. Also, since the U^V values are really very close to 0, we could extract the potential conclusion that model is able to forecast efficiently even when the real motion data vary significantly (e.g. different operators).

Finally, Figure 6 presents an example of trajectory forecasting for GV_2 . More specifically, it shows that when asking to all the State-Space models for the $RWRIST_x$ of $G_{2,1}$ to forecast an unknown observation of the same gesture, we conclude to two results. The simulated values of the $RWRIST_x$ of $G_{2,1}$ for time t , when providing it with real observations until $t - 1$ (starting from $t = 3$), as well as the real observations between t and the end of the sequence, for every time t , is minimum from the first time-stamp (thus high similarity).

4.4 Sensitivity analysis

As mentioned, the GOM depicts all the possible relationships that take place during the process of the performance of a gesture. Following the GOM, the next steps are the estimation of the model, its dynamic simulation and its sensitivity analysis. All those steps lead to checking the model's structure, forecasting ability and its reaction to shocks of its variables respectively.

The sensitivity analysis of the simulated GOM follows two steps. During the first step, all the simulated values of the model are being estimated, after an artificial shock is provoked for the first two frames. In the second step, all the simulated values that came up after the disturbance, are being compared to the simulated values before it (baseline). For example, in Figure 7 the simulated values of $RWRIST_X$ are depicted before (red color) and after (blue color) the disturbance on the values of $RWRIST_Y$ by 80%. The disturbance on the simulated variables of $RWRIST_X$ is observed for ten frames in total, eight more frames than the duration of the initial shock. A similar behavior is also observed for $RWRIST_Y$. The quick adaptation of the models after the application of the artificial shock is observed, which also confirms the low sensitivity of the models to external disturbances.

5 Discussion

The proposed method for human movement representation on multivariate time series has been used for recognition of professional gestures and forecasting of their trajectories. A comparison has been done between the recognition results of our hybrid approach and the standard continuous HMMs. In general, with both approaches, the best recognition accuracy is achieved for the GV_1 . This can be explained by the beneficial *inter*- and *intra*- class variation of this vocabulary. The gestures are sufficiently discrete, while the different repetitions performed by one operator are sufficiently similar. Nevertheless, we observe an improvement on the recognition accuracy for micro-gestures, when the confidence control of the HMM^{SS} is applied for micro-movements, e.g. assembling small pieces, while the performance of HMMs is satisfactory for macro-movements.

The second-best results are given for the GV_2 . Even though these gestures are simpler, and do not require any particular dexterity, less good results in recognition accuracy in comparison to the GV_1 are expected mostly because of the high *intra* class variation due to multiple users. Although, they followed a protocol each person had significant variations in the way he/she performed the commands. For both datasets a slight improvement of results has been achieved.

As explained in section 3.2, the biggest difference of the GV_3 in comparison to the other two gesture vocabularies is the low *inter*-class variation since the gestures are similar between them. In 3 out of 4 gestures common gestural patterns are presented: the glass master if controlling the pipe with the left hand, is manipulating the glass with the right while sitting etc. These common gestural patterns generate the low *intra*-class variation. This low variation can be due to the high level of expert's dexterity, the use of a predefined physical set up (metallic construction) that places his body and gestures in a spatial framework (situated gestures) and the use of professional tools that also reduces potential freedom in gesture performances. The low *intra*- class variation is also underlined by the comparison of the RMSE values for different repetitions of the same gesture performed by the same person. The HMMs are thus expected to provide the less good results among the four datasets, for the GV_3 , since this method may struggle in managing low *inter* class variation. An important similarity between classes is expected to augment the uncertainty in the maximum likelihood probabilities given by the HMMs. This hypothesis can be confirmed through the current recognition results based on HMMs. However, it can be clearly noticed that HMM^{SS} had the most beneficial impact on the recognition

accuracy of the GV_3 . A conclusion can be thus formulated that the proposed methodology permits the improvement of the gesture recognition results to a significant level.

The recognition results of all the three gestural vocabularies using machine learning methods were also compared to those when using 3DCNNs as a DL method for gesture recognition. In all of the three experiments, the HMM^{SS} method outperformed, and especially in GV_1 achieved +3% higher $f - score$ and accuracy compared to the 3DCNNs. In the gestural vocabularies GV_1 and GV_3 the HMM method, even if it was not combined with the SS method achieved slightly higher $f - score$ results than the 3DCNNs.

As far as the GV_4 is concerned, our current approach of continuous HMMs and SS outperforms our previous one that used K-Means and discrete HMMs (Table 7). More precisely, an improvement of at least +12% is observed on the mean $f - score$, together with an improvement of at least +10% at the total accuracy.

As far as the ability of the models to effectively simulate the professional gestures is concerned, the graphical depiction of the simulated values of the models together with the real motion data can lead to encouraging conclusions. Initially, the simulated values follow very well the real ones for the whole gesture. Especially the results on GV_1 are quite promising because the pose estimation had some fails because of the top-mounted camera. Nevertheless, the changes or discontinuities on the motion data didn't affect the simulation ability of the models. With the regard to the forecasting ability of the models, it is obvious that if the model follows the trajectory from the very beginning then its forecasting ability is maximized, which is the case in the Figure 6. Moreover, the evaluation of the forecasting ability of the models using the coefficient of Theil is also encouraging, thus opening a possibility for an efficient forecasting of motion trajectories. In parallel the sensitivity analysis applied to equations variables proves forecasting's ability of the model to react rapidly to shocks and to provide a solid prediction of motion trajectories.

6 Conclusion and future work

In this paper, a Gesture Operation Model is proposed to describe how the body parts cooperate to perform a professional gesture. Several assumptions are formulated that determine the dynamic relationship between the body entities within the execution of the human movement. The model is based on the State-Space statistical representation and a simultaneous equations system for all the body entities is generated, which is composed by a set of first-order differential equations. The coefficients of the equation system are estimated using the Maximum Likelihood Estimation (MLE) and its simulation generates a tolerance of the spatial variance of the movement over time. The scientific evidence of the GOM is evaluated through its ability to improve the recognition accuracy of gestural time series that are modeled using continuous Hidden Markov Models. Four datasets have been created for this experiment, corresponding to professional gestures from industrial real-life scenarios. The proposed approach overperformed the recognition accuracy of the HMMs by approximately +2% for two datasets while a more significant improvement of +10% has been achieved for the third dataset with strongly situated professional gestures. Furthermore, the approach has been compared with an End-to-End 3D Convolutional Neural Network approach and the mean $f - score$ of the proposed method is significantly higher than the DL, varying approximately from +1.57% to +2.9% better performance depending on the dataset. A second comparison is done by using a previously recorded industrial dataset from a Human-Robot Collaboration. The proposed approach gives approximately +13% for the mean $f - score$ and +12% for total accuracy, compared to our previous hybrid K-Means and discrete HMMs approach.

Moreover, the system is simulated through the solution of its equations. Its forecasting ability has been evaluated by comparing the similarity between the real and the simulated motion data, using at least two real observations to initialize the system, as well as by measuring the Theil's inequality coefficient and its decompositions. This paper opened a potential for investigating simultaneous real-time probabilistic gesture and action recognition, as well as forecasting of human motion trajectories for accident prevention and very early detection of the human intention. Therefore, our future work will be focused on extending the proposed methodology for real-time recognition and enhancing the Gesture Operational Model to include kinetic parameters as well. Finally, there will be a continuous enrichment of the datasets by adding new users and more iterations.

7 References

- [1] Duprey, Naaim, Moissenet, Begon, Cheze. "Kinematic models of the upper limb joints for multibody kinematics optimisation: An overview". *Journal of Biomechanics*, Elsevier, 2017, 62, pp. 87-94, [ff10.1016/j.jbiomech.2016.12.005](https://doi.org/10.1016/j.jbiomech.2016.12.005). [ffhal-01635103](https://doi.org/10.1016/j.jbiomech.2016.12.005).
- [2] Zalmai, Kaeslin, Bruderer, Neff, Loeliger. "Gesture recognition from magnetic field measurements using a bank of linear state space models and local likelihood filtering", *IEEE 40th International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Brisbane, Australia, April 19-24, 2015.
- [3] Lech, and Kostek, "Hand gesture recognition supported by fuzzy rules and Kalman filters", *Int. J. Intelligent Information and Database Systems*, Vol. 6, No. 5, 2012 407.
- [4] Pedersoli, Benini, Adami, and Leonardi. "XKin: an open source framework for hand pose and gesture recognition using Kinect", © Springer-Verlag Berlin Heidelberg 2014, *Vis Comput* DOI [10.1007/s00371-014-0921-x](https://doi.org/10.1007/s00371-014-0921-x).
- [5] Aggarwal, and Cai. "Human Motion Analysis: A Review", *Computer Vision and Image Understanding*, Volume 73, Issue 3, 1999, Pages 428-440, ISSN 1077-3142, <https://doi.org/10.1006/cviu.1998.0744>.
- [6] Sideridis, Zacharakis, Tzagkaraki, and Papadopoulou. "GestureKeeper: Gesture Recognition for Controlling Devices in IoT Environments", arXiv:1903.06643v1 [cs.HC] 15 Mar 2019.
- [7] Ruiduo, and Sudeep. "Gesture Recognition using Hidden Markov Models from Fragmented Observations", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, 1. 766- 773. [10.1109/CVPR.2006.126](https://doi.org/10.1109/CVPR.2006.126).
- [8] Li, Koping, Schmitz, and Grzegorzek. "Real-Time Gesture Recognition using a Particle Filtering Approach", *ICPRAM 2017*.
- [9] Jirak, Barros, and Wermter. "Dynamic Gesture Recognition Using Echo State Networks", *ESANN 2015 proceedings*, European Symposium on Artificial Neural Networks. Portillo-Rodríguez, Sandoval-Gonzalez, Ruffaldi, Leonardi, Avizzano, and Bergamasco. "Real-Time Gesture Recognition, Evaluation and Feed-Forward Correction of a Multimodal Tai-Chi Platform", *HAID*, 2008.
- [10] Vaitkevičius, Taroza, Blažauskas, Damaševičius, Maskeliūnas, Woźniak, "Recognition of American Sign Language Gestures in a Virtual Reality Using Leap Motion", *Applied Sciences*, 2019.
- [11] Psarrou, Gong, and Walter. "Recognition of human gestures and behaviour based on motion trajectories", *Image and vision computing*, 2001.
- [12] P. Kolesnik, M.M. Wanderley, "Implementation of the Discrete Hidden Markov Model in Max/MSP Environment", In Proc. of the FLAIRS, 2005, 68-73.
- [13] Bettens, and Todoroff. "Real-time dtw-based gesture recognition external object for max/msp and puredata", *Proceedings of the SMC 2009 Conference*, 30, 35, 2009.
- [14] Françoise. "Motion-sound mapping by demonstration", Ph.D. Thesis, Pierre and Marie Curie University, France, 2015.
- [15] Caramiaux, Montecchio, Tanaka, and Bevilacqua. "Adaptive Gesture Recognition with Variation Estimation for Interactive Systems", *ACM TiiS*, 4, 4, 2015.
- [16] Bevilacqua, Zamborlin, Sypniewski, Schnell, Guédy, and Rasamimanana. "Continuous real time gesture following and recognition", *Proceedings of the 8th International Conference on Gesture in Embodied Communication and Human-Computer Interaction*, Bielefeld, Germany, 2009.

- [17] Bevilacqua, Muller, and Schnell. "MnM: a Max/MSP mapping toolbox", *Proceedings of the NIME'05*, Vancouver, Canada, 2005.
- [18] Bevilacqua, Guédy, Schnell, Fléty, and Leroy. "Wireless sensor interface and gesture-follower for music pedagogy", *Proceedings of the NIME'07*, New York, NY, 2007, 124-129.
- [19] Bobick, and Wilson. "A state-based approach to the representation and recognition of gesture", *IEEE TPAMI*, 19, 12, 1997, 1325-1337.
- [20] Françoise. "Gesture-Sound Mapping by Demonstration in Interactive Music Systems", *Proceedings of the 21st ACM MM'13*, Barcelona, Spain, France, 2013, 1051-1054.
- [21] Françoise, Caramiaux, and Bevilacqua. "A Hierarchical Approach for the Design of Gesture-to-Sound Mappings", *Proceedings of the CMC Conference*, Copenhagen, Denmark. 2012.
- [22] Volioti, Manitsaris, Katsouli, Manitsaris. "x2Gesture: how machines could learn expressive gesture variations of expert musicians", 2016.
- [23] Williamson, and Murray-Smith, "Audio Feedback for Gesture Recognition", 2003.
- [24] Caramiaux. "Optimising the Unexpected: Computational Design Approach in Expressive Gestural Interaction", *Proceedings of the CHI Workshop on Principles, Techniques and Perspectives on Optimization and HCI*, Seoul, Korea, 2015.
- [25] Yang, Park, Lee. "Gesture Spotting and Recognition for Human-Robot Interaction", *IEEE Transactions on Robotics*, Vol. 23, No. 2, April 2007.
- [26] Moutarde, Coupeté, and Manitsaris. "Multi-users online recognition of technical gestures for natural human-robot collaboration in manufacturing", Feb 2018.
- [27] Mathe, Mitsou, Spyrou, and Mylonas. "Hand Gesture Recognition using a Convolutional Neural Network", 2018.
- [28] Oyedotun, Khashman. "Deep learning in vision-based static hand gesture recognition", *Neural Computing and Applications*, 28(12), pp.3941-3951, 2017.
- [29] Molchanov, Gupta, Kim, and Kautz. "Hand gesture recognition with 3D convolutional neural networks", *Proceedings of IEEE conference on computer vision and pattern recognition workshops*. (pp. 1-7), 2015.
- [30] Camgoz, Hadfield, Koller, Bowden. "Using convolutional 3d neural networks for user-independent continuous gesture recognition", *In Pattern Recognition (ICPR)*, 2016, 23rd International Conference on (pp. 49-54). IEEE, , 2016, December.
- [31] Li, Yu, Wu, Su, Ji. "Feature learning based on SAEPCA network for human gesture recognition in RGBD images", *Neurocomputing*, 151, pp.565-573, 2015.
- [32] Devineau, Xi, Moutarde, and Yang. "Deep Learning for Hand Gesture Recognition on Skeletal Data", *13th IEEE Conference on Automatic Face and Gesture Recognition (FG'2018)*, May 2018, Xi'an, China. [ff10.1109/FG.2018.00025](https://doi.org/10.1109/FG.2018.00025)[ff](https://doi.org/10.1109/FG.2018.00025ff). [ffhal-01737771](https://doi.org/10.1109/FG.2018.00025ff).
- [33] Volioti, Manitsaris, Hemery, Charisis, Hadjileontiadis, Hadjidimitriou, Katsouli, Moutarde, and Manitsaris. "A Natural User Interface for Gestural Expression and Emotional Elicitation to Access the Musical Intangible Cultural Heritage", *Journal on Computing and Cultural Heritage*. 11. 10.1145/3127324, 2018.
- [34] Coupeté, Moutarde, Manitsaris, "Gesture recognition using a depth camera for human robot collaboration on assembly line", *ScienceDirect 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences*, AHFE 2015.
- [35] Zatsiorsky, "Biomechanics in Sport: Performance Enhancement and Injury Prevention", *International Olympic Committee*, January 2000.
- [36] Jurafsky, Martin. *Speech and Language Processing*. Copyright c 2019. 2019.

- [37] Cao, Hidalgo, Simon, Wei, and Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.", 2018, 1812.08008, arXiv.
- [38] Baum. "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes", *Proceedings of the Third Symposium on Inequalities*, New York, USA, 1972.
- [39] Liporace. "Maximum likelihood estimation for multivariate observations of Markov sources", *IEEE Trans. Inform. Theory*, IT-28, 729-734, 1982.
- [40] Juang. "Maximum likelihood estimation for mixture multivariate stochastic observation of Markov chains", *AT&T Tech. Journal*, 1235-1249, 1985.
- [41] Juang, Levinson, and Sondhi. "Mixture autoregressive hidden Markov models for speech signals", *IEEE Trans. Acoust., Speech Signal Processing*, 33(6), 307-309, 1985.
- [42] Alani. "Modèles de Markov Cachés - Théorie et techniques de base", *ESIEE*, France, 1994.
- [43] Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, 77(2), 257-285, 1989.
- [44] Dempster. "Maximum Likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society, Series B (Methodological)*, 39, 1-38, 1977.
- [45] Holmes. "Kalman filtering for maximum likelihood estimation given corrupted observations", University of Washington.
- [46] Bakis. "Continuous speech recognition via centisecond acoustic states", *The Journal of the Acoustical Society of America*, New York, 59(1), 97, 1976.
- [47] Przemyslaw Dymarski. "Hidden Markov Models: Theory and Applications", 2011/4/19.
- [48] Makridakis, Wheelwright, Hyndman, "Forecasting: Methods and Applications (3rd edition)", New York: Wiley, 1997.
- [49] Yun, "Human activity recognition and prediction", Switzerland, Springer, 2016.
- [50] <https://www.rolandberger.com/en/Point-of-View/Automotive-manufacturing-requires-human-innovation.html>, Accessed: 08/09/2019.
- [51] Jun, and Wu. "A General Theory for Jackknife Variance Estimation." *The Annals of Statistics*, vol. 17, no. 3, 1989, pp. 1176–1197. JSTOR, www.jstor.org/stable/2241717.
- [52] Manitsaris, Glushkova, Katsouli, Manitsaris, Volioti. "Modelling gestural know-how in pottery based on state-space estimation and system dynamic simulation", *6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences*, AHFE 2015.
- [53] Coupeté, Moutarde, Manitsaris. "Multi-users online recognition of technical gestures for natural human–robot collaboration in manufacturing" *Autonomous Robots* 43, no. 6 (2019): 1309-1325.
- [54] Tran, Bourdev, Fergus, Torresani, and Paluri. "Learning spatiotemporal features with 3d convolutional networks" in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
- [55] Kingma, and Ba, "Adam: a method for stochastic optimization," *arXiv:1412.6980*, 2014.
- [56] Chollet et al., "Keras," <https://keras.io>, 2015.
- [57] Shahroudy, Liu, Ng, and Wang. "Ntu rgb+ d: A large scale dataset for 3d human activity analysis" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.
- [58] Yan, Xiong, and Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition" in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

- [59] Simonyan, and Zisserman. “Two- stream convolutional networks for action recognition in videos,” in *Advances in neural information processing systems*, 2014, pp. 568–576.
- [60] Dimitropoulos, Barmoutis, Kitsikidis, and Grammalidis. "Classification of Multidimensional Time-Evolving Data Using Histograms of Grassmannian Points". *IEEE Transactions on Circuits and Systems for Video Technology*. PP. 1-1. 10.1109/TCSVT.2016.2631719, 2016.
- [61] Cao, Hidalgo, Simon, Wei, Sheikh. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”, 2018.

8 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

9 Funding

The research leading to these results has received funding by the EU Horizon 2020 Research and Innovation Programme under grant agreement No. 820767, CoLLaboratE project, and No. 822336, Mingei project.

10 Acknowledgments

We acknowledge the Arçelik factory as well as Cerfav, the European Centre for Research and Training in Glass Arts - CRT for glass workers and national innovation centre for crafts and especially Jean-Pierre Mateus, master glassblower and trainer for contributing to the motion capturing.

Part of this research was supported of the Chair ‘PSA Peugeot Citroën Robotics and Virtual Reality’, led by MINES ParisTech and supported by PEUGEOT S.A. The partners of the Chair cannot be held accountable for the content of this paper, which engages the authors’ responsibility only.

11 Data Availability Statement

The datasets generated for this study are anonymous and by the time of the manuscript submission, the authors do not have authorization from the industries to publish them. Negotiation is being done with the companies and organizations involved, to make the data publicly available.