frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 13 August 2020
doi: 10.3389/frobt.2020.00080

Human Movement Representation on Multivariate Time Series for Recognition of Professional Gestures and Forecasting Their Trajectories

*Sotiris Manitsaris**, *Gavriela Senteri*, *Dimitrios Makrygiannis and Alina Glushkova*

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

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, for example, gestural commands to machines and professional actions with or without the use of tools. Moreover, the *inter*class 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 equation 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 for both 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. The hybrid approach State-Space and HMMs outperforms standard continuous HMMs and a 3DCNN-based end-to-end deep architecture.

Keywords: state-space representation, differential equations, movement modeling, hidden Markov models, gesture recognition, forecasting, motion trajectory

OPEN ACCESS

Edited by:
Shalabh Gupta,
University of Connecticut,
United States

Reviewed by:
Patrick Henaff,
Université de Lorraine, France
Pedro Neto,
University of Coimbra, Portugal

***Correspondence:**
Sotiris Manitsaris
sotiris.manitsaris@mines-paristech.fr

Specialty section:
This article was submitted to
Sensor Fusion and Machine
Perception,
a section of the journal
Frontiers in Robotics and AI

Received: *20 December 2019*
Accepted: *08 May 2020*
Published: *13 August 2020*

Citation:
Manitsaris S, Senteri G,
Makrygiannis D and Glushkova A
(2020) Human Movement
Representation on Multivariate Time
Series for Recognition of Professional
Gestures and Forecasting Their
Trajectories. Front. Robot. AI 7:80.
doi: 10.3389/frobt.2020.00080

INTRODUCTION

Human motion analysis and recognition are 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 (Borl, 2018). 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, for example, 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, for example, 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, for example, variation in luminosity, various workspace with different geometries, camera in motion to record a person moving in space, and low availability of real experts. Additionally, the community of actions and gesture recognition deals with challenges that are related to intraclass and *inter*class variations (Fu, 2016). Frequent are the cases where a professional task involves gestures that have very similar spatiotemporal characteristics (*low interclass variation*) together with very important differences in the way different humans perform the same gesture (*high intraclass 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 of a set of first-order differential equations. The coefficients of the equation system are estimated using the Maximum Likelihood Estimation (MLE) method, 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 simulated motion data using two real observations for initializing the models as well as by measuring the Theil 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 are 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 (Cao et al., 2019).

State of the Art presents a state of the art on human motion modeling, representation, and recognition. In *Methodology*, the whole methodology analysis, modeling, and recognition are presented, whereas in *Evaluation*, the different approaches in the evaluation of the ability of the models to simulate a gesture and forecast its trajectories are analyzed. In the same section, the accuracy of the proposed method is also presented. Finally, in *Discussion* and *Conclusion and Future Work*, a discussion and the future work and perspectives of the proposed methodology are described.

STATE OF THE ART

In professional environments, a movement can be defined as the displacement of the human body in space, whereas gesture is a form of non-verbal communication for interacting with a machine or manipulating a tool or object. In industry, professional gestures define the routine of workers. The non-physical interaction of the workers with collaborative machines is relying on an external perception layer that gets input from the ambient environment using, among others, human motion sensors, to understand their movement and adapt their behavior according to the humans. Whether machine or deep learning can be used to create an AI-based perception layer for collaborative machines. Machine learning has demonstrated an important number of applications in action and gesture recognition by using whether a probabilistic modeling of the phenomenon, and optionally a representation of it, or a template matching that makes use of a temporal rescaling of the input signal according to the reference. Finally, architectures of deep neural networks have recently seen a considerable number of applications with a high accuracy and precision.

Movement Modeling and Representation

Each body articulation is strongly affected by the movement of other articulations. 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. (2017) 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 (Zatsiorsky, 2000). The representation of the human movement with physical or statistical models provides with a simplified mathematical formalization of the phenomenon and approximates reality, e.g. through simulation and forecasting. State-Space (SS) is a statistical modeling that allows to stochastically represent a human movement through a reasoning over time that makes use of internal states. An 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. (2015) used linear SS models and provided an algorithm based on local likelihood for detecting and inferring gesture causing magnetic field variations. Lech and Kostek (2012) 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. (2016) 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 dynamic system (LDS). Here, multidimensional evolving data were considered as a cloud of points (instead of a single point) on the Grassmann manifold, and codebook is created to represent each multidimensional signal as a histogram of Grassmannian points, which is not always the case for professional gestures.

Machine Learning for Gesture Recognition

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 a template-based method that has been widely used for measuring the similarity between motion data. DTW makes it possible to find the optimal global alignment between two sequences. Bevilacqua et al. (2005), Bevilacqua et al. (2007), and Bevilacqua et al. (2009) 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 (Bobick and Wilson, 1997). 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, Psarrou et al. (2002) used the Conditional Density Propagation algorithm to perform gesture recognition and made sure that they will not get probabilities for only one model per time stamp. The experiments resulted to a relatively good accuracy for the time period conducted.

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 HMMs. Pedersoli et al. adopted this method (Pedersoli et al., 2014) to recognize in real-time static hand poses and dynamic hand gestures of American Sign Language. Sideridis et al. (2019) created a gesture recognition system for everyday gestures recorded with inertia measurement units, based on Fast Nearest Neighbors, and Support Vector Machine methods, whereas Yang and Sarkar (2006) chose to use an extension of HMMs. Vaitkevičius et al. (2019) used also HMMs with the same purpose, gesture recognition, for the creation of virtual reality installations, as well as Williamson and Murray-Smith (2002), who 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 et al. (2007) 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 (Li et al., 2017). HMMs seem to be a solid approach, allowing to achieve 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. (2015) 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 and size) are estimated between the template and the performed gesture (Caramiaux, 2015; Caramiaux et al., 2015).

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 SS method for human movement representation and modeling and to use this representation as the extra control layer to improve gesture recognition results.

Deep Architectures for Action Recognition

Deep Learning (DL) is another approach with increasing scientific evidence in action and gesture classification. Mathe et al. (2018) presented results on hand gesture recognition with the use of a Convolutional Neural Network (CNN), which is trained on Discrete Fourier Transform images that resulted from raw sensor readings. In Oyedotun and Khashman (2017), 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 Molchanov et al. (2015) to detect hand movements of drivers, and in Camgoz et al. (2016) continuously recognize gesture classes from the Continuous Gesture Dataset (ConGD), which is the larger user-independent dataset. A two-stage approach is presented in Li et al. (2015), which combines feature learning from RGB-D using CNNs with Principal Component Analysis (PCA) for selecting the final features. Devineau et al. (2018) used a CNN model and tested its performance on classifying sequential humans' tasks using hand-skeletal data as input. Shahroudy et al. (2016), 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. (2018) 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 (Tran et al., 2015) or two stream networks (Simonyan and Zisserman, 2014). 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 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, and craftsmanship. 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 for both 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. To underline the advantages of the method proposed we also compare its performance with results obtained with a recent DL-based end-to-end architecture.

Our Previous Work

Manitsaris et al. (2015b) 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 Manitsaris et al. (2015b), as well as in Volioti et al. (2016), 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 with position data from various real-life situations that have more classes and users than in previous work.

METHODOLOGY

Overview

The motion capturing of the operators in their workplace is a major task. A number of professional gestural vocabularies are 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 *intrajoint association*, the *interlimb synergies*, the *intralimb mediation*, and the *transitioning* over time. Following the theory of the SS modeling, the GOM is translated into a *simultaneous equation system* that is composed of 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 equation 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 equation 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 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

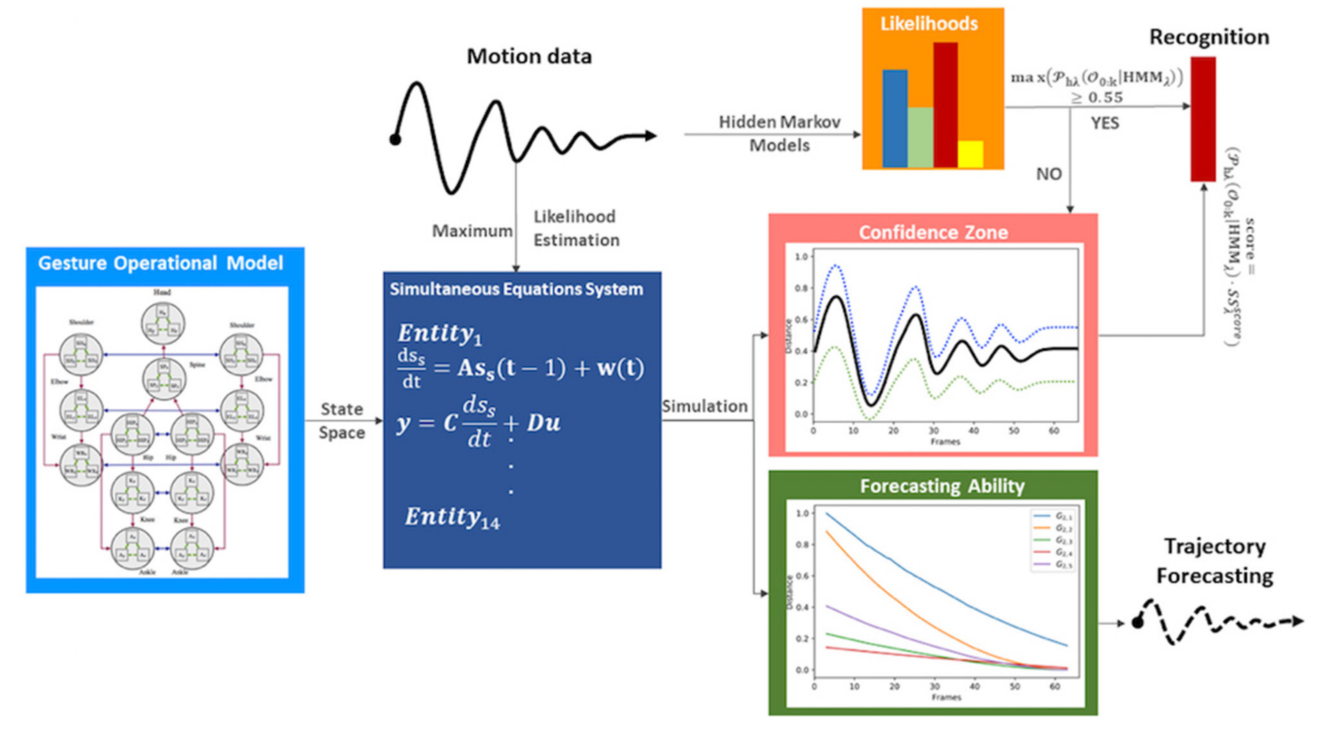

Motion data
Gesture Operational Model
Head
Shoulder
Spine
Elbow
Wrist
Hip
Knee
Ankle
State Space
Maximum Likelihood Estimation
Simultaneous Equations System
$Entity_1$
$\frac{ds_s}{dt} = As_s(t-1) + w(t)$
$y = C\frac{ds_s}{dt} + Du$
$Entity_{14}$
Simulation
Hidden Markov Models
Likelihoods
$\max(\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_\lambda)) \geq 0.55$
YES
NO
Recognition
score = $(\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_\lambda) \cdot SS_\lambda^{score})$
Confidence Zone
Distance
Frames
Forecasting Ability
$G_{2,1}$
$G_{2,2}$
$G_{2,3}$
$G_{2,4}$
$G_{2,5}$
Trajectory Forecasting

FIGURE 1 | Methodology pipeline.

by giving as input at least two time-stamp values from the real motion data.

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 (**Table 2**).

The first gesture vocabulary (GV_1) includes four gestures where the operator takes the electronic card from one box and 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 sides. However, the operator has a certain degree of variation in the way of executing the tasks, because the gestures are ample involving the whole body. Moreover, in order to avoid self-occlusions and scene occlusions, the camera is mounted on the top, which is not necessarily the optimal camera location for pose estimation algorithms, for example, 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 gestural vocabulary proposes gestural commands for controlling an AGV. This dataset GV_2 contains five gestures involving mostly the arm and forearm. $G_{2,1}$ initiates the communication with the AGV, by shaking the palm, whereas $G_{2,2}$ 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, whereas $G_{2,5}$ speeds down the AGV by rolling the right hand away from the hips with a distance of around 20/30 cm. 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. 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 performs various manipulations of the glass by using tools, such as pliers. 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/her 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) used in this paper is related to a real-life Human–Robot Collaboration scenario that has been recorded

[1] https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/output.mdwhwhoch

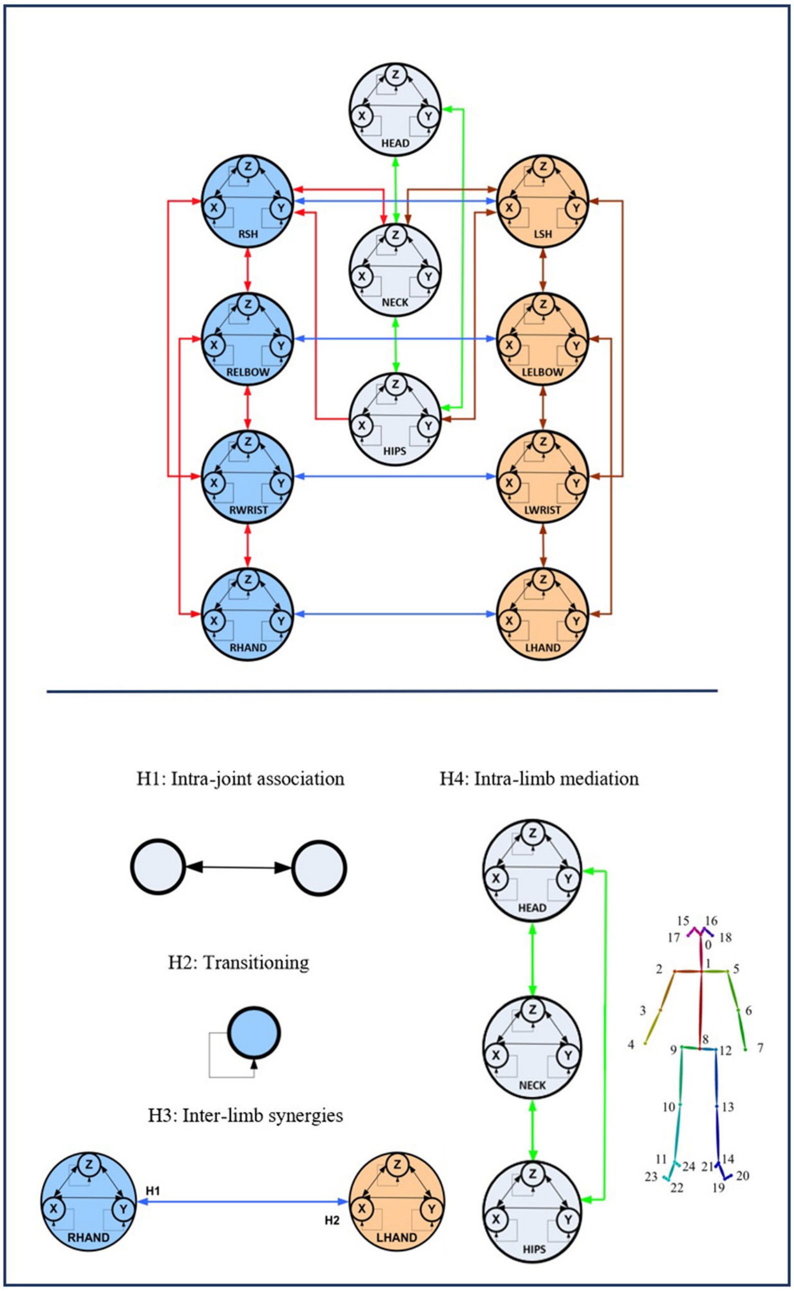


FIGURE 2 | The full-body assumptions of the Gesture Operational Model (GOM) are depicted in the figure. Some relationships happen to be bidirectional, while others not. The relationships of the human body are governed by four different assumptions, intra-joint association, transitioning, inter-limb synergies, and intra-limb mediation. On the down-right of the image, the mapping on body is presented. The numbers in the GOM model, represent the corresponding body part of the joints representation from OpenPose framework. The idea of the GOM was based on a previous article of the authors (Manitsaris et al., 2015b)[1].

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 from the left claw, and 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 subtask. 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, for example, when moving for grabbing the tool, and low values, for example, 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 low spatial and temporal variations can be accepted. The glass blower performs the gestures with a high repeatability from one repetition to another and successfully reproduces the object with exactly the same specifications, for example, size and diameter. 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. Because the dual-arm robot Motoman SDA20 has been used, the operator, depending on whether he/she is left- or right-handed, has various possibilities for grabbing the parts from the robot and the tools.

In GV_1, although all the gestures are performed by a single user, the different positions of the operator in space in each gesture make 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 *intraclass* variance point of view, the Root-Mean-Square Error (RMSE) is used to evaluate the datasets. The RMSE allows the measurement of the difference between two times series, and it is defined as shown in Equation (1).

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

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, because this dataset consists of gestures performed by six different users (**Table 1**). The *RMSE* for GV_3 appears to have low *intraclass* variation, as expected, because the gestures are performed by an expert, who is able to repeat them in a very precise way.

TABLE 1 | RMSE between the iterations of the real data of GV_2 and GV_3.

***GV*₂**	$\overline{G_{2,1}}$	$\overline{G_{2,2}}$	$\overline{G_{2,3}}$	$\overline{G_{2,4}}$	$\overline{G_{2,5}}$
RMSE	0.0565	0.0523	0.0556	0.0330	0.0407
***GV*₃**	$\overline{G_{3,1}}$	$\overline{G_{3,2}}$	$\overline{G_{3,3}}$	$\overline{G_{3,4}}$	
RMSE	0.0265	0.0302	0.0489	0.0461	

Pose Estimation and Feature Extraction

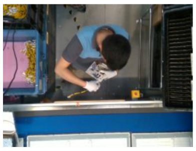
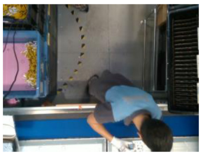
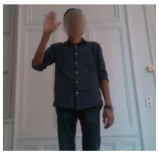
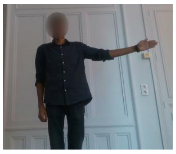
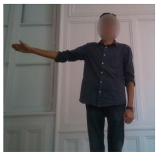
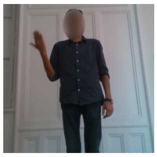
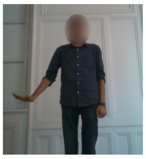
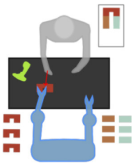
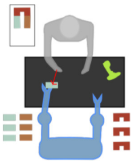
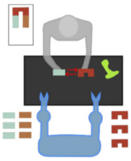
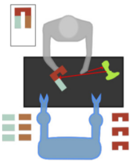
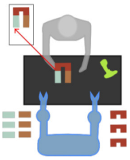
After the motion capturing and recording of the data, each image sequence of the three first datasets is imported to the OpenPose framework, which detects body keypoints on the RGB image and extracts a skeletal model together with the 2D positions of each body joint (Cao et al., 2019) (**Figure 2**). 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 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 (Manitsaris et al., 2015a). 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 does not use any skeletization 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, whereas for GV_2 and GV_3, the 2D positions of the head, neck and shoulder, elbows, wrists, and hands have been used, as they were proven to give optimal recognition results. With regard to GV_4, 3D hand positions are used.

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, for example, from specific postures (small tolerance for spatial variance) to ample movements (high tolerance for spatial variance). For each phase of the movement,

TABLE 2 | Gesture vocabulary of TV assembling dataset, AGV commands dataset, Glassblowing and Human-Robot Collaboration dataset, respectively.

***GV_1* -TV assembly**				
$G_{1,1}$: Take the card from the left side box 	$G_{1,2}$: Take the wire from the right-side box 	$G_{1,3}$: Connect the wire with the card 	$G_{1,4}$: Place the card on the TV chassis 	
***GV_2*-AGV commands**				
$G_{2,1}$: Hello 	$G_{2,2}$: Left 	$G_{2,3}$: Right 	$G_{2,4}$: Speed up 	$G_{2,5}$: Speed down 
***GV_3*-Glassblowing**				
$G_{3,1}$: Fix details with pliers 	$G_{3,2}$: Tighten base of glass 	$G_{3,3}$: Make shape with paper 	$G_{3,4}$: Fix shape 	
***GV_4*-Human-robot collaboration**				
$G_{4,1}$: Take a motor hose part in the robot right claw 	$G_{4,2}$: Take a motor hose part in the robot left claw 	$G_{4,3}$: Join two parts of the motor hose 	$G_{4,4}$: Screw 	$G_{4,5}$: Put the final motor hose in a box 

each body entity, for example, articulation or segment, moves in a multidimensional space over time. When considering the 2D motion descriptors of the movement, two mutually dependent variables represent the entity, for example, 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, whereas some entities might "work together" to execute an effective gesture, for example, when an operator assembles 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, for example, when a potter shapes the clay with both hands.

The above observations on situated body movements can be translated into a functional model, which we define here as the 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 "intrajoint association," "transitioning," "intralimb synergies," and "intralimb mediation" contribute at a certain level to the production of the gesture. As far as the *intralimb mediation* is concerned, it can be decomposed into the "interjoint serial mediation" and the "interjoint non-serial mediation." The proposed model works perfectly for all three dimensions (X, Y, and Z), but for reasons of simplicity, 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.

Intrajoint Association

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

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 .

Interlimb Synergies

It is assumed that some body entities work together to achieve certain motion trajectories, for example, hands when assembling two parts, defined here as *interlimb synergies*.

Intralimb Mediation

Interjoint serial mediation

It is assumed that a body entity may depend on its neighboring entities to which it is directly connected to; for example, 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 *interjoint serial mediation*.

Interjoint non-serial mediation

It is assumed that each body entity depends on non-neighboring entities of the same limb; for example, 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 shoulders, respectively. Accordingly, LELBOW and RELBOW represent the left and right elbows; LWRIST and RWRIST, the left and right wrists; and LHAND and RHAND, the left and right hands. 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 as follows:

$$\begin{aligned} Entity_{1,X}(t) = {} & Entity_{1,Y}(t-1) + Entity_{1,X}(t-1) + \\ & + Entity_{1,X}(t-2) + Entity_{2,X}(t-1) \\ & + Entity_{3,X}(t-1) \end{aligned} \tag{2}$$

Simultaneous Equation System

The simultaneous equation system concatenates the dynamics of an Nth-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.

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. An 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 an 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:

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

$$y = C\frac{ds_s}{dt} + Du \tag{4}$$

where Equation (3) is the *state* equation, which is a first-order Markov process; Equation (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 (w(t)) are also added in both the state and output equations. After the experiments presented in this work were performed, it was observed that Gaussian disturbances did not 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 an SS form and j the dimension of each Entity-* according to the GOM structured, as follows:

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

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

where α_i are the coefficients that need to be estimated. For simplicity, the inter-joint non-serial mediation is not used in the specific example. In Equation (6), $Entity_X(t-2)$ is subtracted by $Entity_X(t-1)$, indicating the difference between successive levels of dimensions, for example, positions on the 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 now the form of a second-order autoregressive (AR) model. An AR model predicts future behavior on the basis of past behavior. The order of the

AR model is adapted in each case according to the data characteristics and the experiments. During the performance of the experiments, the use of an AR 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 equation system is based on Equations (5) and (6), which consists of two 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 variables with 64 state variables that contain two endogenous variables for $t-1$ and $t-2$.

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

$$\begin{aligned} RWRIST_X(t) &= \alpha_1 RWRIST_X(t-1) - \alpha_2 RWRIST_X(t-2) + \\ &+ \alpha_3 RWRIST_Y(t-1) + \alpha_4 LWRIST_X(t-1) \end{aligned} \tag{7}$$

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

Computing the Coefficients of the Equations

The coefficients of the simultaneous equation system are computed using the MLE method via Kalman filtering (Holmes, 2016). Let us consider a gesture $G_{j\in\mathbb{N}}$ of a gesture vocabulary $GV_{i\in[1,3]}$ and an observation $\mathcal{O}_{0:k} = \{o_1, \ldots, o_k\}_{k\in\mathbb{N}}$, 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:

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

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

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

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

$$\begin{aligned} \log\mathrm{L}(\Psi|\mathcal{O}_{0:t-1}) &= -\frac{k}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{k}\log\left|F_t^{t-1}\right| - \\ &-\frac{1}{2}\sum_{t=1}^{k}\frac{(o_t - \tilde{o}_t^{t-1})^2}{F_t^{t-1}} \end{aligned} \tag{10}$$

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.

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 (Bakis, 1976). In our case, each professional gesture is associated to an HMM, whereas 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 is between the states $\mathcal{Q}\left(s_h, s_h'\right)$, where $s_h, s_h' \in \mathcal{S}_h$ are given in the transition matrix $\mathcal{Q} = \left[\mathcal{Q}\left(s_h, s_h'\right)\right]$. A hidden sequence of states where $\mathcal{S}_{h0:k} = \{s_{h1}, \ldots, s_{hk}\}_{k\in\mathbb{N}}$, where $s_{hk} \in \mathcal{S}_h$ is also considered. A given sequence of hidden states $S_{h0:k}$ is supposed to generate a sequence of observation vectors $\mathcal{O}_{o:k}$. We assume that the vectors o_k depends only on the state s_{hk}. From now on, the likelihood that the observation o is the result of the state s_h will be defined as $\mathcal{P}_h(o|s_h)$. 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 $GV_{j\in[1,3]} = \{HMM_i\}_{i\in\mathbb{N}}$, where $HMM_i = (\varrho_i, \mathcal{Q}_i, \mathcal{P}_{hi})_{i\in\mathbb{N}}$ are the parameters of the model and ϱ_i is the initial state probability. Thus, the recognition becomes an issue of solving three specific problems: *evaluation*, *recognition*, and *learning* (Dymarski, 2011). Each one of those problems was solved with the use of the algorithms, Viterbi (Rabiner, 1989), Baum's "forward" (Baum, 1972), and Baum–Welch, respectively (Dempster et al., 1977).

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 us 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 $i \in [1, 3]$ generates the likelihood $\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_\lambda)$. 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):

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

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

Then, for every SS_{λ} of GV_i, the likelihood $\mathcal{P}'_{h\lambda}\left(\mathcal{O}_{0:k}|HMM_{\lambda}^{SS}\right)$ is computed as follows:

$$\mathcal{P}'_{h\lambda}\left(\mathcal{O}_{0:k}|HMM_{\lambda}^{SS}\right) = \mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_{\lambda}) \cdot SS_{\lambda}^{score} \tag{12}$$

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

$$\mathcal{R}_{GV_i}(\mathcal{O}_{0:k}) = \begin{cases} \max_1^j\left(\mathcal{P}_{hi}(\mathcal{O}_{0:k}|HMM_i)\right), & \max\left(\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_{\lambda})\right) \geq 0.55 \\ \max_1^j\left(\mathcal{P}'_{h\lambda}\left(\mathcal{O}_{0:k}|HMM_{\lambda}^{SS}\right)\right), & \max\left(\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_{\lambda})\right) < 0.55 \end{cases} \tag{13}$$

EVALUATION

The evaluation of the accuracy and performance of the method follows an "all-shots" approach for the training of the HMMs and a "one-shot" approach for estimating the coefficients of the SS models.

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

STATISTICAL SIGNIFICANCE AND SIMULATION OF THE MODELS

In order to evaluate the significance of the assumptions concerning the body part 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 0 and 1. 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, whereas others cannot, and can remain constant. In the GOM below, the equation of $G_{2,1}$ for $RWRIST_X$ is as follows, starting from Equation (2).

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

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, whereas 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 axis and the left wrist remains still through the performance of the gesture, leading to the result that there is no intralimb mediation in this specific gesture.

In the 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.

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

$$\begin{aligned} RWRIST_Y(t) &= a_{12}RWRIST_X(t-1) + a_{13}RWRIST_Y(t-1) - \\ &\quad -a_{14}RWRIST_Y(t-2) + a_{15}LWRIST_Y(t-1) \\ &= \overbrace{-3.9907}^{0.00} RWRIST_X(t-1) + \\ &\quad +\overbrace{0.5003}^{0.00} RWRIST_Y(t-1) - \end{aligned}$$

$$
\begin{aligned}
&-\overbrace{(-0.0818)}^{0.616}RWRIST_{Y}(t-2)+\\
&+\overbrace{(-0.0927)}^{0.00}LWRIST_{Y}(t-1)
\end{aligned}
\tag{16}
$$

In this gesture, the operator moves his/her right wrist toward his/her 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 two previous time-period values 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, whereas the left wrist complements the movement by slowly rolling the metal pipe.

$$
\begin{aligned}
RWRIST_X\,(t) \;=\;& a_{12}RSH_X\,(t-1)+a_{13}RELBOW_X(t-1)+\\
&+a_{14}RWRIST_Y\,(t-1)+a_{15}LWRIST_X(t-1)+\\
&+a_{16}\,RWRIST_X\,(t-1)-a_{17}RWRIST_X(t-2)\\
=\;& \left(\overbrace{-0.0778}^{0.562}\right)RSH_X\,(t-1)+\overbrace{1.1126}^{0.00}\,RELBOW_X(t-1)+\\
&+\left(\overbrace{-0.4757}^{0.00}\right)\,RWRIST_Y\,(t-1)+\overbrace{0.3423}^{0.00}\times\\
&\times LWRIST_X\,(t-1)+\overbrace{0.4585}^{0.00}\,RWRIST_X(t-1)+\\
&-\overbrace{0.4604}^{0.00}\,RWRIST_X(t-2)
\end{aligned}
\tag{17}
$$

$$
\begin{aligned}
RWRIST_Y\,(t) \;=\;& a_{12}RSH_Y\,(t-1)+a_{13}RELBOW_Y(t-1)+\\
&+a_{14}RWRIST_X(t-1)+a_{15}LWRIST_Y(t-1)+\\
&+a_{16}\,RWRIST_Y\,(t-1)-a_{17}RWRIST_Y\,(t-2)\\
=\;& \overbrace{0.290}^{0.117}\,RSH_Y\,(t-1)+\overbrace{0.3678}^{0.00}RELBOW_Y(t-1)+\\
&+\left(\overbrace{-1.0912}^{0.00}\right)\,RWRIST_X(t-1)+\left(\overbrace{-0.1602}^{0.045}\right)\times\\
&\times LWRIST_Y(t-1)+\overbrace{1.1298}^{0.00}\,RWRIST_Y(t-1)+\\
&-\overbrace{(-0.1679)}^{0.00}\,RWRIST_Y(t-2)
\end{aligned}
\tag{18}
$$

$$
\begin{aligned}
LWRIST_X\,(t) \;=\;& a_{12}LSH_X\,(t-1)+a_{13}LELBOW_X(t-1)+\\
&+a_{14}\,LWRIST_Y\,(t-1)+a_{15}RWRIST_X(t-1)+\\
&+a_{16}\,LWRIST_X\,(t-1)-a_{17}LWRIST_X\,(t-2)\\
=\;& \overbrace{0.3668}^{0.00}\,LSH_X\,(t-1)+\overbrace{0.11180}^{0.007}\,LELBOW_X(t-1)+\\
&+\overbrace{0.9589}^{0.00}\,LWRIST_Y\,(t-1)+\left(\overbrace{-0.0126}^{0.339}\right)\times\\
&\times RWRIST_X\,(t-1)+\overbrace{1.1111}^{0.00}LWRIST_X(t-1)+\\
&-\overbrace{(-0.1398)}^{0.052}\,LWRIST_X(t-2)
\end{aligned}
\tag{19}
$$

$$
\begin{aligned}
LWRIST_Y\,(t) \;=\;& a_{12}LSH_Y\,(t-1)+a_{13}LELBOW_Y(t-1)+\\
&+a_{14}LWRIST_X\,(t-1)+a_{15}RWRIST_Y(t-1)+\\
&+a_{16}LWRIST_Y\,(t-1)-a_{17}LWRIST_Y\,(t-2)\\
=\;& \overbrace{0.9313}^{0.00}\,LSH_Y\,(t-1)+\overbrace{0.5433}^{0.00}\,LELBOW_Y\,(t-1)\\
&+\overbrace{0.1144}^{0.272}\,LWRIST_X(t-1)+\overbrace{0.0162}^{0.356}\,RWRIST_Y\,(t-1)+\\
&+\overbrace{1.0463}^{0.0}\,LWRIST_Y(t-1)-\overbrace{(-0.1095)}^{0.124}\,LWRIST_Y(t-2)
\end{aligned}
\tag{20}
$$

In the equations presented above, all coefficients appear to be statistically significant, except from $RSH_X\,(t-1)$ in Equation (17), $LWRIST_X\,(t-1)$ in Equation (18), $LWRIST_X\,(t-1)$ and $RWRIST_X\,(t-2)$ in Equation (19), and $RWRIST_X\,(t-1)$, $RWRIST_Y\,(t-1)$, *and* $LWRIST_X\,(t-1)$ in Equation (20). As a result, the hands of the operator work mostly independently (there appears to be a dependency in the interlimb synergies in Equation 17), whereas all the other assumptions seem to be statistically significant for both X-axis and Y-axis of the right and left wrists.

The simulation of the models is based on the solution of their simultaneous equations system. **Figures 3–5** show examples of the graphical depiction of real observations of motion data together with their simulated values from the SS 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.

Recognition Accuracy 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.

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

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

Precision, *recall*, and *f-score* are calculated for all the gestures that each gestural vocabulary consists of. For a gesture of class *i*, #(*true positives*) represent the number of gestures of class *i* that were recognized correctly, #(*false positives*) represent the number of gestures that did not belong in class *i*, and they were recognized from the algorithm as parts of class *i*. Finally, #(*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*, whereas *recall* represents the rate of iterations of gestures of class *i* that have been recognized as class *i*. A measure that

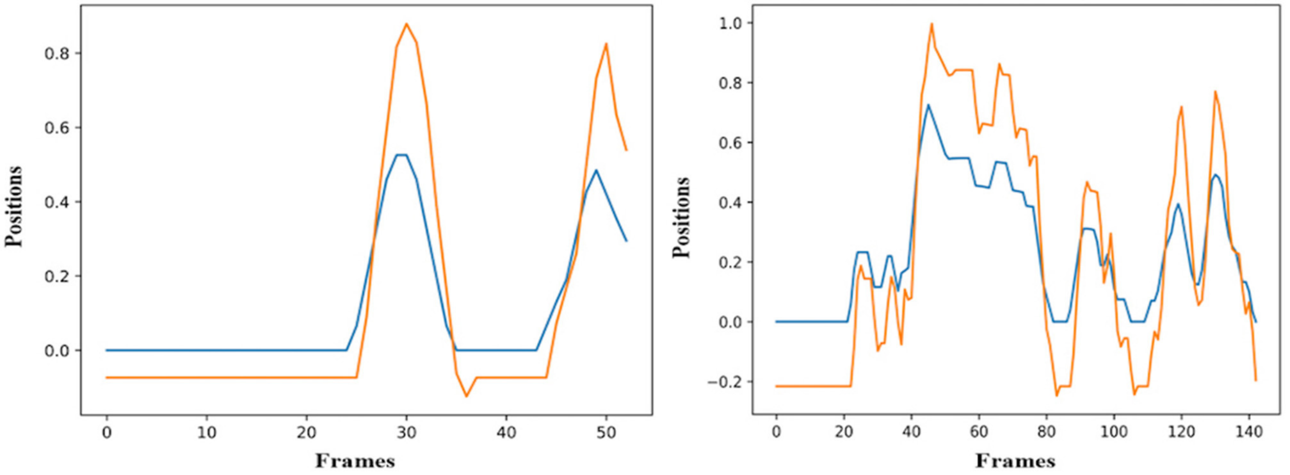


FIGURE 3 | Examples of real motion observations (blue) and simulated values (orange) from the $RHAND_X$ State-Space model of the gesture $G_{1,1}$ (left) and $G_{1,4}$ (right).

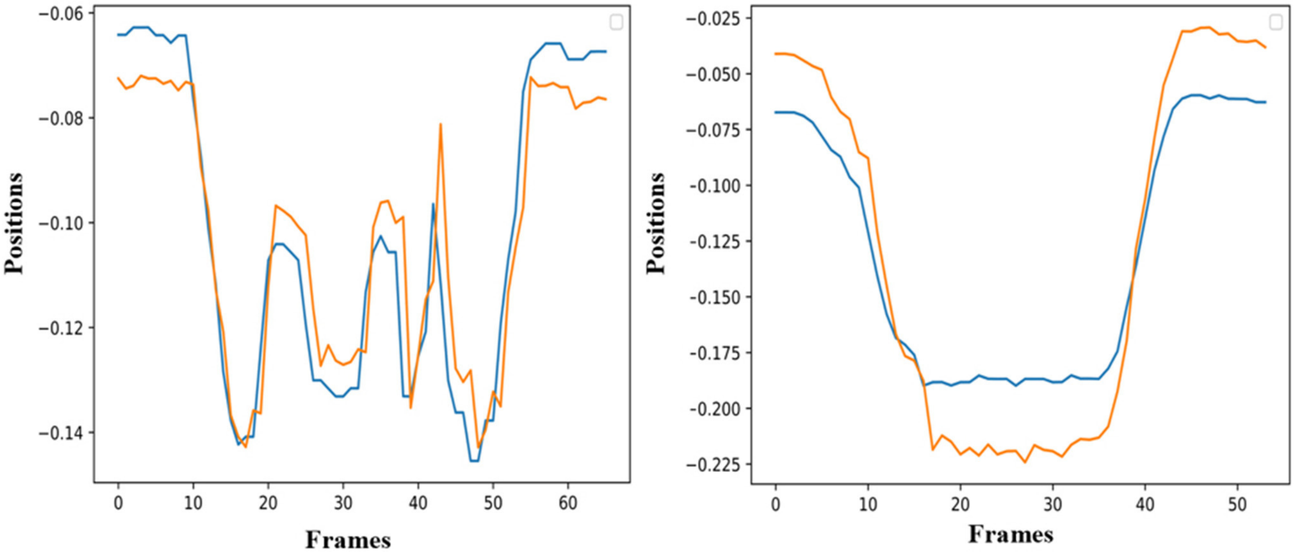


FIGURE 4 | Examples of real motion observations of motion data (blue) and simulated values (orange) from the $RHAND_X$ State-Space model of gesture $G_{2,1}$ (left) and $G_{2,2}$ (right).

combines both precision and recall is the f-score, which is given by Equation (23).

$$f - score = 2\frac{precision{*}recall}{precision + recall} \tag{23}$$

The performance of the algorithms was tested with the four different gestural vocabularies. As presented before, the GV_1 contains four classes, from 44 to 48 repetitions for each. Four hidden states were used for HMM 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 SS models.

It is possible to observe that HMMs provide a recall superior to 90% in three out of four 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 three 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, whereas 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

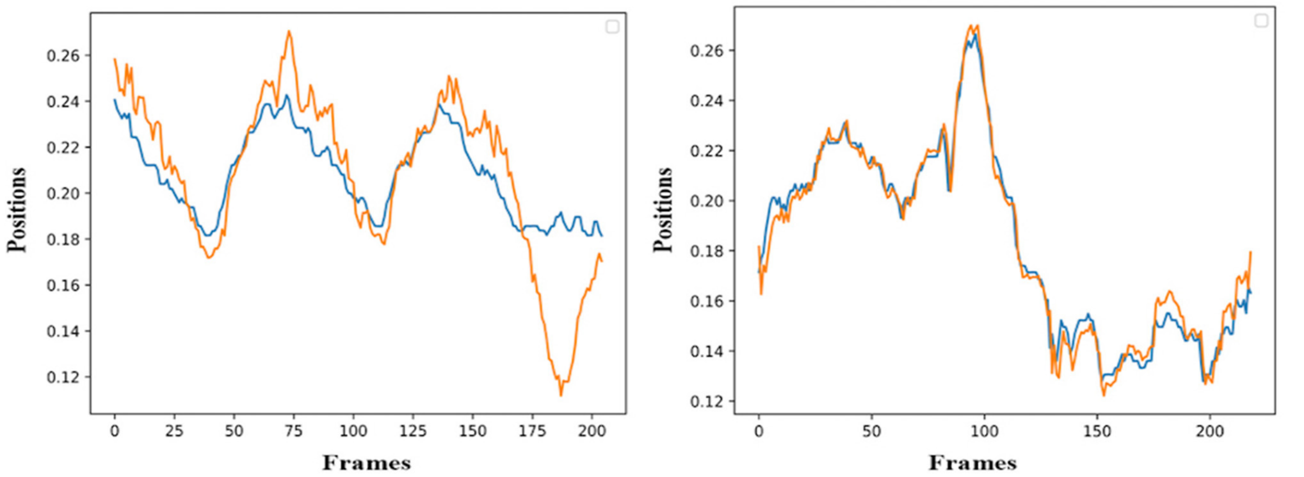


FIGURE 5 | Examples of real motion observations (blue) and simulated values (orange) from the $RHAND_X$ State-Space model of the gesture $G_{3,1}$ (left) and $G_{3,4}$ (right).

TABLE 3 | Confusion matrix using HMM, HMM^{SS}, and $3DCNN$ based on the model of Tran et al. (2015) approaches for GV_1.

	$HMM_{1,1}$	$HMM_{1,2}$	$HMM_{1,3}$	$HMM_{1,4}$	**Recall (%)**
$G_{1,1}$	48	0	0	0	100
$G_{1,2}$	0	44	0	2	95.65
$G_{1,3}$	1	4	36	3	81.81
$G_{1,4}$	0	0	0	46	100
Precision (%)	97.9	91.66	100	90.2	
	$HMM_{1,1}^{SS}$	$HMM_{1,2}^{SS}$	$HMM_{1,3}^{SS}$	$HMM_{1,4}^{SS}$	**Recall (%)**
$G_{1,1}$	47	0	0	1	97.91
$G_{1,2}$	1	45	0	0	97.82
$G_{1,3}$	0	1	43	0	97.72
$G_{1,4}$	0	4	0	42	91.3
Precision (%)	97.91	90	100	97.67	
	$3DCNN_{1,1}$	$3DCNN_{1,2}$	$3DCNN_{1,3}$	$3DCNN_{1,4}$	**Recall (%)**
$G_{1,1}$	48	0	0	0	100
$G_{1,2}$	0	43	0	5	89.5
$G_{1,3}$	0	0	44	0	100
$G_{1,4}$	0	7	0	39	84.7
Precision (%)	100	86	100	88.6	

workplace. The *precision* of $HMM_{1,4}^{SS}$ has been also positively impacted by the SS augmenting the precision 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 five classes and 16 repetitions of each gesture, and one to 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 4**, *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, whereas left to right topology was used for the rest of the gestures. *Precision* appears improved for every model, whereas *recall* is decreased for $G_{2,2}$ and $G_{2,5}$ (**Table 4**).

GV_3 consists of four different gestures with 35, 34, 21, and 27 repetitions, respectively. 5 to 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 of the machine learning methods, with f-score results improved by +12% (**Table 6**).

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 ∼2%, while the most important contribution is observed for the GV_3. The HMM^{SS} allows 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 DL end-to-end 3D CNN has been used to classify the gestures of the three first vocabularies described in *Industrial Datasets and Gesture Vocabularies*. More precisely, a $3DCNN$ has been initially trained on spatiotemporal features from a medium-sized UCF-101 video

TABLE 4 | Confusion matrix using HMM, HMM^{SS}, and $3DCNN$ based on the model of Tran et al. (2015) approach for GV_2.

	$HMM_{2,1}$	$HMM_{2,2}$	$HMM_{2,3}$	$HMM_{2,4}$	$HMM_{2,5}$	**Recall (%)**
$G_{2,1}$	14	1	0	1	0	87.5
$G_{2,2}$	3	13	0	0	0	81.25
$G_{2,3}$	0	0	16	0	0	100
$G_{2,4}$	5	0	0	11	0	68.75
$G_{2,5}$	2	0	0	2	12	75
Precision (%)	58.3	92.8	100	78.5	100	
	$HMM^{SS}_{2,1}$	$HMM^{SS}_{2,2}$	$HMM^{SS}_{2,3}$	$HMM^{SS}_{2,4}$	$HMM^{SS}_{2,5}$	**Recall (%)**
$G_{2,1}$	16	0	0	0	0	100
$G_{2,2}$	4	12	0	0	0	75
$G_{2,3}$	0	0	16	0	0	100
$G_{2,4}$	2	0	0	14	0	87.5
$G_{2,5}$	3	0	0	3	10	62.5
Precision (%)	64	100	100	82.3	100	
	$3DCNN_{2,1}$	$3DCNN_{2,2}$	$3DCNN_{2,3}$	$3DCNN_{2,4}$	$3DCNN_{2,5}$	**Recall (%)**
$G_{2,1}$	16	0	0	0	0	100
$G_{2,2}$	0	16	0	0	0	100
$G_{2,3}$	0	0	16	0	0	100
$G_{2,4}$	0	0	0	12	4	75
$G_{2,5}$	8	0	0	0	8	50
Precision (%)	66.6	100	100	100	66.66	

TABLE 5 | Confusion matrix using HMM, HMM^{SS}, and the Tran et al. (2015) $3DCNN$ approach for GV_3.

	$HMM_{3,1}$	$HMM_{3,2}$	$HMM_{3,3}$	$HMM_{3,4}$	**Recall (%)**
$G_{3,1}$	31	2	1	1	88.57
$G_{3,2}$	0	33	1	0	97.05
$G_{3,3}$	2	2	16	1	76.19
$G_{3,4}$	0	0	0	27	100
Precision (%)	93.93	89.18	88.88	93.1	
	$HMM^{SS}_{3,1}$	$HMM^{SS}_{3,2}$	$HMM^{SS}_{3,3}$	$HMM^{SS}_{3,4}$	**Recall (%)**
$G_{3,1}$	31	2	1	1	88.57
$G_{3,2}$	0	33	1	0	97.05
$G_{3,3}$	1	1	17	2	80.95
$G_{3,4}$	0	0	0	27	100
Precision (%)	96.87	91.66	89.47	90	
	$3DCNN_{3,1}$	$3DCNN_{3,2}$	$3DCNN_{3,3}$	$3DCNN_{3,4}$	**Recall (%)**
$G_{3,1}$	35	0	0	0	100
$G_{3,2}$	0	27	7	0	79.4
$G_{3,3}$	2	2	17	0	80.9
$G_{3,4}$	0	0	0	27	100
Precision (%)	94.5	93.1	70.8	100	

TABLE 6 | Confusion matrix using HMM and HMM^{SS} approach for GV_4.

	$HMM_{4,1}$	$HMM_{4,2}$	$HMM_{4,3}$	$HMM_{4,4}$	$HMM_{4,5}$	**Recall (%)**
$G_{4,1}$	42	1	0	0	1	95.4
$G_{4,2}$	3	86	0	0	1	95.5
$G_{4,3}$	0	0	75	2	12	84.2
$G_{4,4}$	1	0	0	43	0	97.7
$G_{4,5}$	2	0	3	0	75	93.7
Precision (%)	87.5	98.8	96.15	95.5	86.2	
	$HMM^{SS}_{4,1}$	$HMM^{SS}_{4,2}$	$HMM^{SS}_{4,3}$	$HMM^{SS}_{4,4}$	$HMM^{SS}_{4,5}$	**Recall (%)**
$G_{4,1}$	42	1	0	0	1	95.5
$G_{4,2}$	3	86	0	0	1	95.5
$G_{4,3}$	0	0	87	0	2	89.8
$G_{4,4}$	5	2	0	37	0	84
$G_{4,5}$	1	0	5	0	74	92.5
Precision (%)	82.3	96.6	94.5	100	94.8	

dataset, and the pretrained weights have been used to fine-tune the model on small-sized datasets including images of operators performing customized gestures in industrial environments.

The architecture of the network is based on the one proposed in Tran et al. (2015) with four convolution and two pooling layers, one fully connected layer, and a softmax loss layer to predict action labels. It has been trained from scratch on the UCF-101[2] video dataset, using batch size of 32 clips and the Adam optimizer (Kingma and Ba, 2014) for 100 epochs, with the Keras DL framework (Chollet, 2015). The entire network was frozen, and only four last layers were fine-tuned on customized datasets by backpropagation.

[2] https://www.crcv.ucf.edu/data/UCF101.php

TABLE 7 | Comparison of mean *f-scores* and final accuracies of each GV for HMM and HMM^{SS} approach.

Mean f-score	**Datasets**			
	$GV_1\%$	$GV_2\%$	$GV_3\%$	$GV_4\%$
HMM	94.34	83.1	90.64	92.1
HMM^{SS}	96.21	85	91.57	92.29
$k-\text{means}+HMM$	-	-	-	80
$3DCNN$	93.4	84	90	-
Total accuracy	**Datasets**			
	$GV_1\%$	$GV_2\%$	$GV_3\%$	$GV_4\%$
HMM	94.56	82.5	91.45	92.5
HMM^{SS}	96.19	85	92.3	93.94
$k-\text{means}+HMM$				82
$3DCNN$	93.5	87	89	

For the datasets GV_1, GV_2 and GV_3 also the 3DCNN results are presented, based on the model of Tran et al. (2015). For GV_4 the results are compared to those presented in Coupeté et al. (2019) using a k-means and HMM approach, with 25 clusters and discrete HMMs with 12 hidden states.

The comparison of recognition accuracy results between $HMMs$, HMM^{SS}, and $3DCNN$ is shown in **Tables 3**, **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 **Table 3**. However, in total, the HMM^{SS} outperforms the other two methods, reaching a total accuracy of 96.19% (**Table 7**). In the second dataset (GV_2), $3DCNN$ does not achieve a satisfying recognition result for the $G_{2,5}$ (66%) in comparison with other methods that reach 100% (**Table 4**); and in total, HMM^{SS} still performs the best as it is possible to observe in **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.

Forecasting Ability for Motion Trajectories

For the evaluation of the ability of the four SS models that are used to explain the assumptions of the two-entity 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 (Wheelwright et al., 1997). Thus, $U^{\mathrm{B}} + U^{\mathrm{V}} + U^{\mathrm{C}} = 1$. The 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 0 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 0 and U^{C} is close to 1. The computed coefficients as shown in **Table 8** and result to a sufficient forecasting performance of the simulated model, and the error results reinforce this conclusion. Also, because 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).

TABLE 8 | Theil inequality coefficient, root mean squared error, for one example of the X coordinate of the right wrist per dataset.

Gestures	**Theil Inequality *U***	**Bias proportion *U*^B**	**Variance proportion *U*^V**	**Covariance proportion *U*^C**	**RMSE**
$G_{1,1}$	0.018388	0.009178	0.081456	0.909366	0.028904
$G_{2,1}$	0.0000373	0	0.017247	0.983653	0.007461
$G_{3,1}$	0.0000161	0	0.008713	1.041715	0.003277
$G_{4,1}$	0.010059	0	0.039551	0.960449	0.018053

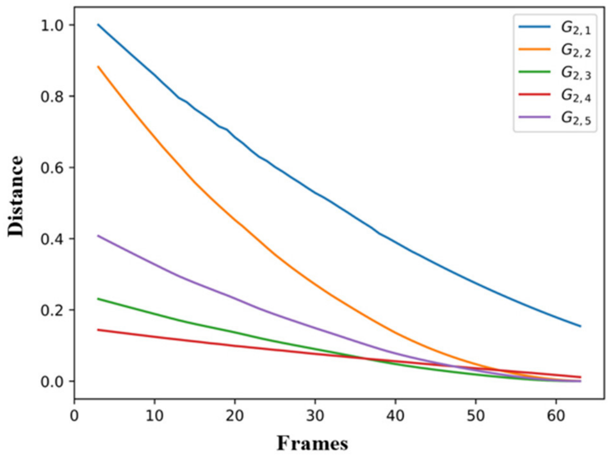


FIGURE 6 | Forecasting performance of all the SS models of GV_2 on the variable $RWRIST_X$ based on $G_{2,4}$.

Finally, **Figure 6** presents an example of trajectory forecasting for GV_2. More specifically, the forecasting performance of all the SS models of GV$_2$ on the variable $RWRIST_X$ is presented, when an unknown observation with data from the $G_{2,1}$ is provided to them. The similarity or distance metric from the DTW is plotted on **Figure 6** taking as input for every time t: 1) the simulated values of the $RWRIST_X$ on $G_{2,4}$ when providing it with real observations until $t = 1$ (starting from $t = 3$), and 2) the real observations between t and the end of the sequence. The distance becomes minimum (high similarity) from the very first time-stamp for the SS model of $G_{2,4}$.

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.

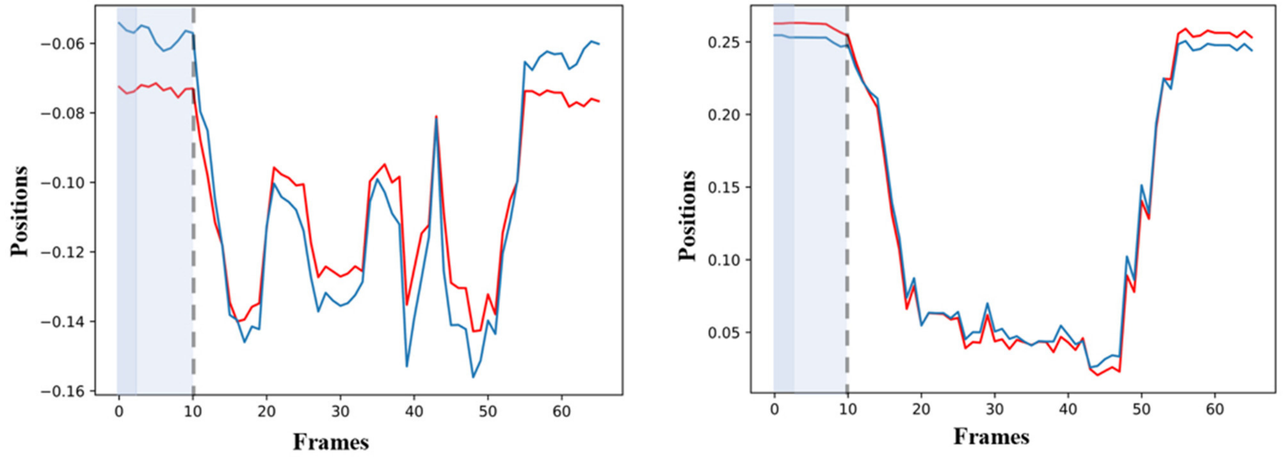


FIGURE 7 | Left: Diagram of the simulated forecasted values of $RWRIST_X$ before the disturbance (red) and simulated forecasted values of $RWRIST_X$ (blue), after the shock on the values of $RWRIST_Y$ by 80% for two frames. Right: Diagram of the simulated forecasted values of $RWRIST_Y$ before the disturbance (red) and simulated forecasted values of $RWRIST_Y$ after the shock on the values of $RWRIST_X$ by 80% for two frames.

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 with 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 10 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.

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*class and *intra*class variations of this vocabulary. The gestures are sufficiently discrete, whereas 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, for example, assembling small pieces, whereas the performance of *HMM*s 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 with the GV_1 are expected mostly because of the high intraclass 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 *Industrial Datasets and Gesture Vocabularies*, the biggest difference of the GV_3 in comparison with the other two gesture vocabularies is the low *inter*class variation because the gestures are similar between them. In three out of four 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, and so forth. 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 setup (metallic construction) that places his/her 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 less good results among the four datasets, for the GV_3, because this method may struggle in managing low interclass 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 on the basis of *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 with those when using 3*DCNNs* 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 with 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. In particular, 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 did not 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 **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.

CONCLUSION AND FUTURE WORK

In this paper, a Gesture Operational 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 SS statistical representation, and a simultaneous equation system for all the body entities is generated, which is composed of a set of first-order differential equations. The coefficients of the equation system are estimated using the 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 HMMs. 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, whereas 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 3DCNN 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 $\sim$ +13% for the mean f-score and +12% for total accuracy, compared with our previous hybrid k-means and discrete HMM 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 simulated motion data, using at least two real observations to initialize the system, as well as by measuring the Theil 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 GOM to include kinetic parameters as well. Finally, there will be a continuous enrichment of the datasets by adding new users and more iterations.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available. The datasets generated for this study are anonymous and by the time of the article 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.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

SM: Conceptualisation of the methodology and definition of the scientific and industrial needs to address with respect to the human motion analysis, machine learning, and pattern recognition. GS: Contribution to the recording of the dataset, implementation of the experiments, and configuration of the statistical parameters of State-Space. DM: Leading the recording of the datasets, pose estimation and festure selection, developing and integrating the algorithms, and MLE and kalman filtering fine tuning the forecasting ability of the models. AG: Definition of the protocol for the recordings, human factor analysis of the motion data and definition of the vocabularies, and interpretation of the recognition results.

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 Grant No. 822336, Mingei project.

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 Center for Crafts and especially Jean-Pierre Mateus, master glassblower and trainer, for contributing to the motion capturing. Part of this research was supported by 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.

REFERENCES

Bakis, R. (1976). Continuous speech recognition via centisecond acoustic states. *J. Acoust. Soc. Am.* 59:S97. doi: 10.1121/1.2003011

Baum, L. E. (1972). "An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," in *Proceedings of the Third Symposium on Inequalities* (New York, NY).

Bevilacqua, F., Guédy, F., Schnell, N., Fléty, E., and Leroy, N. (2007). "Wireless sensor interface and gesture-follower for music pedagogy," *Proceedings of the NIME'07* (New York, NY), 124–129. doi: 10.1145/1279740.1279762

Bevilacqua, F., Müller, R., and Schnell, N. (2005). "MnM: a Max/MSP mapping toolbox," in *Proceedings of the NIME'05* (Vancouver, BC).

Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., and Rasamimanana, N. (2009). "Continuous real time gesture following and recognition," *Proceedings of the 8th International Conference on Gesture in Embodied Communication and Human-Computer Interaction* (Bielefeld). doi: 10.1007/978-3-642-12553-9_7

Bobick, A. F., and Wilson, A. D. (1997). A state-based approach to the representation and recognition of gesture. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 1325–1337. doi: 10.1109/34.643892

Borl, H. (2018). *Toyota is Bucking the Industrial Automation Trend and Putting Humans Back on the Assembly Line*. Available Online at: https://www.rolandberger.com/en/Point-of-View/Automotive-manufacturing-requires-human-innovation.html (accessed October 08, 2019).

Camgoz, N. C., Hadfield, S., Koller, O., and Bowden, R. (2016). "Using convolutional 3D neural networks for user-independent continuous gesture recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR)* (Cancun), 49–54. doi: 10.1109/ICPR.2016.7899606

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S. E., and Sheikh, Y. A. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* doi: 10.1109/tpami.2019.2929257

Caramiaux, B. (2015). "Optimising the unexpected: computational design approach in expressive gestural interaction," in *Proceedings of the CHI Workshop on Principles, Techniques and Perspectives on Optimization and HCI* (Seoul).

Caramiaux, B., Montecchio, N., Tanaka, A., and Bevilacqua, F. (2015). Adaptive gesture recognition with variation estimation for interactive systems. *ACM TiiS* 4, 1–34. doi: 10.1145/2643204

Coupeté, E., Moutarde, F., and Manitsaris, S. (2019). "Multi-users online recognition of technical gestures for natural humanŰrobot collaboration in manufacturing" in *Autonomous Robots* 43, 1309–1325.

Chollet, F. (2015). *Keras: Deep Learning Library for Theano and Tensorflow*. Available Online at: https://keras.io (accessed November 25, 2019).

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B* 39, 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Devineau, G., Moutarde, F., Xi, W., and Yang, J. (2018). "Deep learning for hand gesture recognition on skeletal data," in *13th IEEE Conference on Automatic Face and Gesture Recognition (FG'2018)* (Xi'An). doi: 10.1109/FG.2018.00025

Dimitropoulos, K., Barmpoutis, P., Kitsikidis, A., and Grammalidis, N. (2016). Classification of multidimensional time-evolving data using histograms of grassmannian points. *IEEE Trans. Circuits Syst. Video Technol.* 28, 892–905. doi: 10.1109/TCSVT.2016.2631719

Duprey, S., Naaim, A., Moissenet, F., Begon, M., and Cheze, L. (2017). Kinematic models of the upper limb joints for multibody kinematics optimisation: an overview. *J. Biomech.* 62, 87–94. doi: 10.1016/j.jbiomech.2016.12.005

Dymarski, P. (2011). *Hidden Markov Models: Theory and Applications* (IntechOpen). doi: 10.5772/601

Fu, Y. (ed.). (2016). *Human Activity Recognition and Prediction*. Switzerland: Springer. doi: 10.1007/978-3-319-27004-3

Holmes, E. E. (2016). *Kalman Filtering for Maximum Likelihood Estimation Given Corrupted Observations*. Seattle: University of Washington.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015* (San Diego, CA: Conference Track Proceedings). Available online at: http://arxiv.org/abs/1412.6980

Lech, M., and Kostek, B. (2012). Hand gesture recognition supported by fuzzy rules and kalman filters. *Int. J. Intell. Inf. Database Syst.* 6, 407–420. doi: 10.1504/IJIIDS.2012.049304

Li, F., Köping, L., Schmitz, S., and Grzegorzek, M. (2017). "Real-time gesture recognition using a particle filtering approach," in *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods. Vol. 1* (Porto: ICPRAM), 394–401. doi: 10.5220/0006189603940401

Li, S. Z., Yu, B., Wu, W., Su, S. Z., and Ji, R. R. (2015). Feature learning based on SAE–PCA network for human gesture recognition in RGBD images. *Neurocomputing* 151, 565–573. doi: 10.1016/j.neucom.2014.06.086

Manitsaris, S., Glushkova, A., Katsouli, E., Manitsaris, A., and Volioti, C. (2015b). "Modelling gestural know-how in pottery based on state-space estimation and system dynamic simulation," in *6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences* (Las Vegas, NV). doi: 10.1016/j.promfg.2015.07.883

Manitsaris, S., Moutarde, F., and Coupeté, E. (2015a). "Gesture recognition using a depth camera for human robot collaboration on assembly line," in *ScienceDirect 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated* Las Vegas, NV: Conferences (AHFE).

Mathe, E., Mitsou, A., Spyrou, E., and Mylonas, P. (2018). "Arm gesture recognition using a convolutional neural network, in *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)* (Zaragoza), 37–42. doi: 10.1109/SMAP.2018.8501886

Molchanov, P., Gupta, S., Kim, K., and Kautz, J. (2015). "Hand gesture recognition with 3D convolutional neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Boston, MA: IEEE), 1–7. doi: 10.1109/CVPRW.2015.7301342

Oyedotun, O. K., and Khashman, A. (2017). Deep learning in vision-based static hand gesture recognition. *Neural Comput. Appl.* 28, 3941–3951. doi: 10.1007/s00521-016-2294-8

Pedersoli, F., Benini, S., Adami, N., and Leonardi, R. (2014). XKin: an open source framework for hand pose and gesture recognition using kinect. *Vis. Comput.* 30, 1107–1122. doi: 10.1007/s00371-014-0921-x

Psarrou, A., Gong, S., and Walter, M. (2002). Recognition of human gestures and behaviour based on motion trajectories. *Image Vis. Comput.* 20, 349–358.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–285. doi: 10.1109/5.18626

Shahroudy, A., Liu, J., Ng, T. T., and Wang, G. (2016). "Ntu rgb+ d: a large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 1010–1019. doi: 10.1109/CVPR.2016.115

Sideridis, V., Zacharakis, A., Tzagkaraki, G., and Papadopouli, M. (2019). "GestureKeeper: gesture recognition for controlling devices in IoT environments," *27th European Signal Processing Conference (EUSIPCO)* (A Coruña: IEEE). doi: 10.23919/EUSIPCO.2019.8903044

Simonyan, K., and Zisserman, A. (2014). "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems, Vol. 1*. (Montreal: MIT Press), 568–576.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference On Computer Vision* (Las Condes), 4489–4497. doi: 10.1109/ICCV.2015.510

Vaitkevičius, A., Taroza, M., BlaŽauskas, T., Damaševičius, R., Maskeliunas, R., and Wozniak, M. (2019). Recognition of American sign language gestures in a virtual reality using leap motion. *Appl. Sci.* 9:445. doi: 10.3390/app9030445

Volioti, C., Manitsaris, S., Katsouli, E., and Manitsaris, A. (2016). "x2Gesture: how machines could learn expressive gesture variations of expert musicians," in *Proceeding of the 16th International Conference on New Interfaces for Musical Expression.*

Wheelwright, S., Makridakis, S., and Hyndman, R. J. (1997). *Forecasting: Methods and Applications.* 3rd edition. New York, NY: Wiley.

Williamson, J., and Murray-Smith, R. (2002). *Audio Feedback for Gesture Recognition* Technical report TR-2002-127, Dept. Computing Science, University of Glasgow.

Yan, S., Xiong, Y., and Lin, D. (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence* New Orleans, LA.

Yang R., and Sarkar, S. (2006). "Gesture recognition using hidden markov models from fragmented observations," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (New York, NY), 766–773. doi: 10.1109/CVPR.2006.126

Yang, H. D., Park, A. Y., and Lee, S. W. (2007). Gesture spotting and recognition for human–robot interaction. *IEEE Trans. Robot*. 23, 256–270. doi: 10.1109/TRO.2006.889491

Zalmai, N., Kaeslin, C., Bruderer, L., Neff, S., and Loeliger, H. A. (2015). "Gesture recognition from magnetic field measurements using a bank of linear state space models and local likelihood filtering," in *IEEE 40th International Conference on Acoustics, Speech and Signal Processing* (Brisbane, QLD: ICASSP), 19–24. doi: 10.1109/ICASSP.2015.7178435

Zatsiorsky, V. (2000). *Biomechanics in Sport: Performance Enhancement and Injury Prevention.* Blackwell Publishing; International Olympic Committee.

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.