Hand tracking and gesture recognition by multiple contactless sensors: a survey

Eleni Theodoridou, Student Member, IEEE[1], Luigi Cinque, Senior Member, IEEE[2], Filippo Mignosi[3], Giuseppe Placidi, Member, IEEE[1], Matteo Polsinelli[1], João Manuel R. S. Tavares, Member, IEEE[4], and Matteo Spezialetti, Member, IEEE[3]

[1]A2VI-Lab, Department of Life, Health & Environmental Sciences, University of L'Aquila, 67100, L'Aquila, Italy. eleni.theodoridou@graduate.univaq.it

[2]Department of Computer Science, Sapienza University, 00198, Rome, Italy

[3]Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila, 67100, L'Aquila, Italy

[4]Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecâanica, Faculdade de Engenharia, Universidade do Porto, 4200-465, Porto, Portugal

***Abstract*—Hand tracking and gesture recognition are fundamental in a multitude of applications. Various sensors have been used for this purpose, however all monocular vision systems face limitations caused by occlusions. Wearable equipment overcome said limitations, although deemed impractical in some cases. Using more than one sensor provides a way to overcome this problem, but necessitates more complicated designs. In this work, we aim to highlight contemporary methods used for hand tracking and gesture recognition by collecting publications of systems developed in the last decade, that employ contactless devices as RGB cameras, IR and depth sensors, along with some preceding pillar works. Additionally, we briefly present common steps, techniques and basic algorithms used during the process of developing modern hand tracking and gesture recognition systems and, finally, we derive the trend for the next future.**

I. Introduction

Hand tracking systems have a wide range of applications in many fields. The potential of using the hand as an input device for Human Computer Interaction (HCI) is one of the commonest goals, as many Virtual Reality (VR) and Augmented Reality (AR) applications benefit from direct contactless interaction. While limb motion tracking has been initially used for athletic performance measurements [1], this was followed by impressive applications on the medical field: hand monitoring can be used in both diagnostic and therapeutic procedures, as well as for patient monitoring and practice during mobility rehabilitation [2], [3], [4], [5]. Reduced hand mobility can be caused by a wide range of neurological and musculoskeletal conditions such as, between others, strokes, Alzheimer's disease and arthritis [6]. Virtual training of surgeons [7] and teleoperation [8] also benefit from detailed, contactless hand monitoring. Numerous sign language recognition systems have also been developed [9], [10], underlining the potential of improving human's quality of life through Computer Vision. In the industrial field, tracking hand movements is vital for robot training and operation applications [11], including remote object manipulation and accident avoidance [12], [13].

However, the mechanical and functional form of the human hand creates difficulties. Fingers have the ability to move fast, achieving velocities of up to 5 m/s while the rotational speed of the wrist can reach up to 300 degrees/s during normal gestures [14]. The five fingers are similar in shape and color, so they are not easily distinguishable and modeled. During monocular vision, frequent occlusions happen when a part of the hand covers another (self-occlusions), or during the interaction with an object. The situation improve when multiple views are implemented.

The requisites for motion capture vary depending on the application it supports, with the system's spatial and temporal resolution and its robustness being critical attributes. When it comes to the sensing approach, contactless systems allow the tracking of free and natural hand movements with the use of affordable equipment; however, they are subjected to issues like occlusions or detection over cluttered background. On the other hand, wearable devices are not affected by occlusions, but impede the user's mobility, additionally to being prone to noisy data. Medical conditions can also impede the use of wearable sensors [15]. In Table I, a brief comparison of contactless and wearable devices' characteristics is presented.

Though wearable devices have certain advantages which could make them preferable, modern social, health and scientific conditions push us towards the development of new ways for user-independent (not affected from the hand laterality, shape, size and residual impairment), non interfering with the free hand movement, and contactless systems. Consequently, it is important to focus on contactless approaches that, additionally to their technical attributes, grant absolute safety. In this work, we chose to focus on systems proper for contactless and non-personalized data acquisition.

The management of occlusions becomes a major challenge when monitoring hand movements remotely. Many techniques are developed aiming at advancing pose recognition, however improving the way of acquiring data is also vital. Multisensor systems are suitable for studying synergic hand movements,

TABLE I: Comparison between wearable and contactless systems: the best characteristics are indicated in Italic.

Characteristic	Wearable devices	contactless devices
Environmental conditions and background	*Almost independent*	May cause issues
Hand part discrimination	*Easy*	Sometimes hard
Spatial resolution	*High*	Sensor-dependent
Occlusions	*No*	Yes
User-dependency	High	*No*
Interferes with free hand movement	Yes	*No*
Safety	Usually Low	*High*

regarding either hand-to-hand or hand-to-object interactions. By having more than one detector, one can choose the best between several different views, while simultaneously improving the system's robustness. The sensors can work continuously, or only the sensor with the better ''view" of the target can be active at a time. Also, one sensor can be used as the indicator of the optimal view. Furthermore, merging data from multiple sources can significantly improve the overall resolution and accuracy.

There exist very interesting surveys covering the field of hand monitoring. In their fundamental work, the authors of [14] produced a detailed study of contactless, vision-based systems that estimate hand poses, analyzing thoroughly methods that attempt to use the hand as an input device for HCI. In [16], the authors focused on glove-based systems and their application in various domains, while also an interesting comparison of tracking glove technologies is provided. More recently, Chen et al. [17] analyzed motion detection methods by depth sensors. This study is not limited to hand motion but focuses on whole body monitoring. Older reviews [18], [19] show a continuity of efforts towards discovering an effective way of tracking hand motion in real time, since the 1980s. To our knowledge, there is no survey dedicated especially to the subject of multi sensor methods, although their use is spreading rapidly.

For these reasons, we chose to focus on a collection of hand tracking and hand gesture recognition systems that utilize multiple contactless sensors. The reviewed papers were collected from Scopus using the queries "Multi sensor hand tracking", "Multi sensor gesture recognition", "Multiple sensor hand tracking", "Multiple sensor gesture recognition" and "Data fusion hand tracking". Also, material was collected from Google Scholar using the same queries together the search words "hand tracking" and at least one of the words "multisensor" and "contactless" (or "touchless" instead). Following this procedure, and after merging the result of different searches and removing duplicate articles, we collected 1707 works.

However, since our queries contain words commonly used in other senses, our collection includes numerous irrelevant articles (a notable example is that the word "hand" in the query also collects papers with the phrase "on the other hand" in their abstract). These cases were excluded after checking the title and, if needed, at the abstract of the articles. After this we had a total of around 400 publications.

Next, we applied the following inclusion and exclusion criteria: Only studies providing semantics information, that is using clear discrimination between hand parts, for the hand as an articulated object were included. In the same way, works about tracking full body movements or arm movements, without providing information of the hand as an articulated object were filtered off. Studies that utilize wearable sensors, as Inertial Measurement Unit (IMU)s, alone or in combination with contactless techniques, or any kind of device or marker put on the hand, were also excluded. Afterwards, through the citations of these publications, more papers meeting the predefined criteria were gathered. Since our goal was to present recent technological and scientific advances, we collected and studied researches published from 2010 and after. The reason for choosing 2010 as our starting point is twofold. First, many consumer sensors have been proposed after that date; Second, the development of Deep Learning (DL) methods, assisted by the expansion of GPU and relative software usage, created a lot of new data analysis approaches for Computer Vision.

Still, it would be counterproductive to exclude some pillar works published before 2010, but having significantly influenced the following years' research. For this reason, the choice of these works was done based on their citation numbers, by using an indicative threshold of 60 citations.

This approach led to the selection of 33 articles: 21 were from Scopus, 23 from Google Scholar, with a common subset of 11 articles. Other 16 papers of this survey were the result of studying citations of already gathered works.

The rest of the paper is organized as follows. In Section II, basic software and hardware tools used for multisensory hand monitoring are presented. Also, basic methodologies are described, as well as a collection of multisensory datasets. In Section III, the collected papers are presented, separated in pillar works, hand tracking and gesture recognition categories, both using homogeneous and heterogeneous sensors, and mainly organized in chronological order. A conclusive discussion and future trends follows in section VI.

II. Systems' Architecture

In this section we aim to provide the reader with the basic aspects of contactless hand monitoring, together with the equipment used and the main software tools. Our first and main distinction regards the objective. We found that hand monitoring systems fall in two major clusters: hand tracking and gesture recognition. In both cases, a fundamental step is the representation of the hand form and position, that is usually achieved by incorporating a hand model. Spatial data regarding 3-dimensional information are collected either with a number of 2-D RGB cameras or by 3-D sensors, as we describe ahead. Software applications are used for calibrating the sensors, and to either fit the spatial data to the model or for Machine Learning techniques, and involve between others, classification algorithms and optimization methods. Last, we

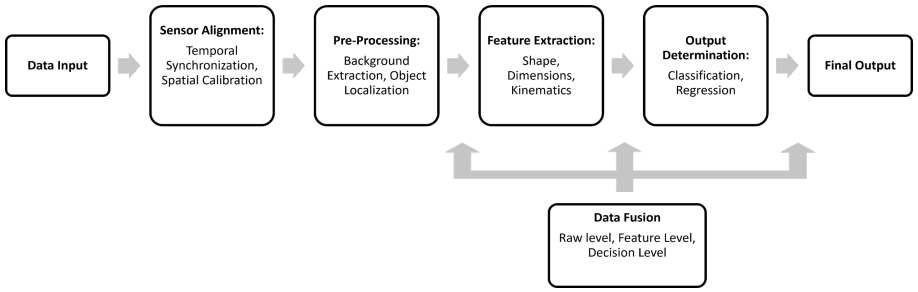

Fig. 1: The general steps of a hand monitoring procedure using multiple sensors. In hand tracking some parameters, i.e. spatial precision and temporal resolution, are more challenging than in gesture recognition.

present hand tracking in a complicated environment where hands interact with objects.

A. Hand monitoring: tracking and gesture recognition

When semantic information is required, according to the application, one needs to perform hand tracking or gesture recognition, by utilizing, respectively, a generative or a discriminative technique.

In hand tracking, the objective is a numerical, time-dependent description of each hand part's position in 3D space with respect to the "world" reference system. These methods are in general computational costly and demand highly detailed image input. In this case, error evaluation is simple if there exist ground truth data. The metric error is frequently calculated as the mean Euclidean distance of the corresponding hand joints from the ground truth.

For gesture recognition, discriminative methods use large sets of real or synthetic training data with ground truth information, to define certain hand poses, as a classification problem. Their computational cost is lower, but they can only recognize certain predefined hand forms. For discriminative systems, data sets with ground truth information are used for evaluating the results, with accuracy being one of the most employed metrics.

It is important to remark that hand tracking also allows for gesture recognition, while the reverse is not true.

B. Hand Models

Hand tracking algorithms aim to analyse spatial data in order to compute the configurations of a hand model and to describe the hand joints' 3D positions and movements, during time.

There exist skeletal hand models, where each hand part is considered as an one-dimensional object. Volume-based models, combined by a set of small geometrical structures have been developed. Volumetric models can be mounted on a skeleton model, for a more realistic result. Linear Blend Skinning (LBS) [20] models, usually adapted on a skeleton, recreate the hand volume and skin surface. The hand parts' volume is in these cases algorithmically estimated or produced by 3D-scanning real hands. In every case, the Degrees of Freedom (DOFs)s described depend on the goal of the application. Identifying and setting certain mechanical and anatomical constraints, boosts the accuracy of the model [14]. Lately, machine learning techniques tend to minimize the need for hand model utilization.

C. Multisensory Hardware

Three basic types of touchless sensors are used for hand monitoring. A small number of RGB cameras provide high spatial and temporal resolution for 2D monitoring, and the possibility for 3D reconstruction, though they need to be calibrated and have a stereo vision of the object to be reconstructed. Stereo RGB (e.g. Point Grey Bumblebee2[1]) systems also rely on the same principle but engineered on a single device. Depth (RGB-D) cameras (e.g. the Kinect sensor[2]) provide images in the visible light together with depth information, acquired by infrared light techniques. However, they may have reduced precision and may be insufficient in outdoor environments [21]. Infrared (IR) stereo systems (e.g. the Leap Motion Controller (LMC) sensor[3]) collect images just in the infrared spectrum while maintaining a good spatial resolution, though in a reduced Field Of View (FOV). In Table II, a brief synopsis of the basic advantages for each kind of system is presented.

TABLE II: 3D Systems Characteristics. The first two columns describe whether the resulting data are affected by the background and the lighting conditions, whereas the last two present some of the sensors' potentials.

Sensor	Background independency	Light invariance	Large FOV	Simple setup
IR Stereo	yes	yes	no	yes
RGB-D	no	no	yes	yes
Multiple RGB	no	no	yes	no
Stereo RGB	no	no	yes	yes

D. Software

As shown in Figure 1, there are several steps that need to be fulfilled for a hand monitoring system.

The spatial calibration and the temporal synchronization of the sensors can be achieved through specialized, usually point set registration algorithms, that utilize a commonly

[1]https://www.flir.com/support/products/bumblebee2-firewire/\#Overview
[2]https://developer.microsoft.com/en-us/windows/kinect/
[3]https://www.ultraleap.com/

visible cloud of points. Then, as a first step of pre-processing, background extraction is performed in order to isolate the hand region, followed by 2-D visual cues extraction, such as color and edges, or 3-D ones, such as silhouettes, salient points and visual hulls.

Some frequently used classifiers are Hidden Markov Models (HMMs) [22], Support Vector Machines (SVMs) [23], while Decision Trees [24] are used both for classification and regression. Deep learning approaches such as Convolutional Neural Networks (CNNs) are also widely developed on the last years. For regression, optimization procedures, frequently based on Particle Swarm Optimization (PSO) [25] techniques are utilized.

Data fusion can be performed right after the data acquisition (Raw Level Fusion (RLF)), after feature extraction (Feature Level Fusion (FLF)) or the outputs of separate procedures can be combined (Decision Level Fusion (DLF)) [26]. Multiple sensors can also be used independently, for broadening the total FOV or for utilizing different sensors' capabilities.

E. Hand-environment interaction

When hands interact either with other hands or with an object, challenging conditions are formed. Occlusions are expected and visual cues such as color and edges cannot be used in cases of close hand-to-hand interactions. Handled objects can also create unusual hand gestures, not included in training datasets. In the last years, new methods have been developed dealing with this issues, since interacting hand tracking is vital in applications that include object handling. Physics simulators and anatomical restrictions can help with determining the plausibility of a solution [27].

The type and quality of data information is crucial for the result of hand monitoring. Especially in Machine Learning approaches, a sufficient training dataset including ground truth information is vital. In Tables III and IV we report a list of commonly used datasets, for multisensor hand tracking and for gesture recognition applications, respectively. We only present publicly available datasets.

III. Computer-vision based hand monitoring strategies

According to our inclusion criteria, 50 papers are presented in this survey. In Table V, a synoptic categorization of these works is shown whereas in Table VI, the basic characteristics of each methodology are displayed. In particular, Table VI is organized by separating first hand tracking and gesture recognition then, inside each category, a separation is done between the usage of sensors of the same kind (homogeneous) or of different kinds (heterogeneous) and, finally, they are sorted by ascending date. In that way, some useful characteristics are easily identified. The last column of the table refers to the research purpose that each article is dedicated on. Specifically, the purposes were grouped in the following large categories: Human-Computer Interaction, Medical application, Robot Manipulation, Signal Language Recognition and Virtual Reality. The purpose of articles that develop methods and techniques without mentioning a special application is referred to as Non Defined.

A. Hand monitoring pillar works before 2010

This section is a collection of researches prior to 2010, which significantly contributed to the development of future systems for the following reasons: 1) they are the pioneering works; 2) they could only rely on RGB sensors, that means 3D information was not implicitly available; 3) they were compelled to design complex deterministic calibration, classification and recognition strategies or to rely on 3D constraints to compensate for the lack of 3D technologies and of powerful computers to train Machine and Deep learning strategies. Since the pillar works are small in number and contain architectures and techniques that have been proved useful at future works, both for external analysis and for hardware development, we chose to analytically describe them. As shown in Table VI, these methods mainly concentrate on gesture recognition and are forced to use multiple homogeneous sensors (RGB), covering views to reduce occlusions or to gather 3D information through stereo-vision.

A.1) Hand tracking

There are three works dedicated on hand tracking [35], [36], [37].

In details, the 1996 work of [35] involves input from 2 cameras and a 3-D model, and is based on maximizing an overlapping function between the model and the images' silhouettes. It uses a conjugated skeletal model, on which a set of geometrical primitives is adjusted. Each segment has its own coordinate system, creating a total of 33 DOFs. This is one of the first works where a synthetic dataset with more than one points of view is utilized.

In [36], occlusions are presented as the main problem of touchless sensors for hand tracking, and multiple sensors are introduced as a way of controlling it. A system composed by n cameras is proposed and its 5-camera example is experimentally tested. Kalman filter (KF) is used for tracking hand position, orientation and shape. Thresholding and a distance transformation define the hand silhouette.

In [37] a multisensor system aiming to hand rehabilitation, the Virtual Glove (VG), is presented. It consists of a set of 4 RGB cameras placed on the vertices of a cubic space surrounding the user's hand. The cameras are calibrated to allow 3D information acquisition from at least two sensors. The redundancy of sensors ensures occlusion minimization. This is one of the first studies that aim especially to hand tele-rehabilitation with the use of multiple sensors, while also introducing the interaction with a foreign object, for the indirect calculation of forces exerted by the fingers. An heterogeneous multisensor version of the VG, including RGB-D together with RGB input, is described in [74].

A.2) Gesture recognition

Four works prior to 2010 regard gesture recognition [54], [55], [56], [57]. In [54], model profiles derived from camera views are compared with a different video sequence, for the hand pose estimation. A volume-based hand model is used,

TABLE III: Multisensor hand tracking public datasets

Dataset	Sequences	Interactions	RGB	IR	Depth	Cited
Tzionas Monocular RGB-D *(http://files.is.tue.mpg.de/dtzionas/Hand-Object-Capture/)*	23	✓	1		1	[28]
Tzionas Multicamera RGB *(http://files.is.tue.mpg.de/dtzionas/Hand-Object-Capture/)*	8	✓	8			[28]
Dexter 1 dataset *(http://handtracker.mpi-inf.mpg.de/projects/handtracker_iccv2013/dexter1.htm)*	7		5		2	[28], [29], [30]

TABLE IV: Multisensor hand gesture public datasets

Dataset	Subject	Hands	Gestures	Samples	RGB	IR	Depth	Cited
Indian Sign Language (ISL) *(https://sites.google.com/site/iitrcsepradeep7/)*	10	1	25	2000	1	1	1	[31], [32]
Microsoft Kinect and LEAP Motion dataset *(http://lttm.dei.unipd.it/downloads/gesture)*	14	1	10	1400	1	1	1	[33]
Nvidia Dynamic Hand Gesture dataset *(https://research.nvidia.com/publication/online-detection-and-classification-dynamic-hand-gestures-recurrent-3d-convolutional)*	20	1	25	1500	1	1	1	[34]
Briareo dataset *(http://imagelab.ing.unimore.it/briareo)*	40	1	12	120	1	1	1	[34]

TABLE V: Contactless multisensor hand monitoring works. Homogeneous are considered the systems composed by two or more identical devices, while heterogeneous systems include different kinds of sensors.

		Tracking	Gesture recogn.	Total
Homogeneous systems	**RGB**	10	6	**16**
	IR	6	10	**16**
	RGB-D		1	**1**
Heterogeneous systems	**RGB & RGB-D**	2		**2**
	IR & RGB-D	6	8	**14**
	IR & RGB		1	**1**
Total		**24**	**26**	**50**

composed by 3D geometrical shapes. Occlusions are reduced by comparing the depth of points in the 3D space.

In [55], input is provided by two RGB cameras functioning as a stereo pair. Only the thumb and the index finger are tracked, aiming to the recognition of simple hand gestures for interactive applications. Edge patterns are detected after contour extraction. An overall detection frequency of 15 Hz is achieved. This work presents interesting pre-processing procedures and a simple calibration method, however it focuses only on the detection of a small number of gestures.

A histogram-based skin-color classifier is applied on [56]. It recovers the hand silhouette, followed by edge detection. This is one of the first works in multisensory hand monitoring that uses a trained classifier with a dynamic hand model.

In a work of 2007, authors of [57] present a real-time two-hand gesture recognition system that deals with the recogition of 4 different gestures, each one defined by 6 DOFs. It is developed for use in HCI applications and can achieve recognition frequencies up to 25 fps. Only two or three fingers are monitored. This work involves both the left and the right human hand, but again its gesture dataset is limited.

B. Hand monitoring from 2010 and on

Starting from 2010, RGB-D and IR sensors appeared and, after few years (around 2015), their cost became affordable for public use (however, RGB sensors have continued to be used until 2016). Two sensor technologies mostly contributed to this step: Microsoft Kinect, for RGB-D sensors, and LMC for IR sensors. These technologies have allowed to: reduce the number of sensors, being no more necessary to use stereo sensors to calculate 3D positions for tracking; have a direct tool to reduce the complexity of the problem, for example by using distance information of RGB-D sensors for background elimination; eliminate, in most cases, the need for calibration; simplify the optimization/classification strategies used for hand monitoring, for example by reducing or eliminating the use of external constraints. All these advantages gave new impulse to the development of ever more effective and efficient hand monitoring systems, listed in Table VI.

B.1) Hand tracking

From 2010 to 2017, a series of works were dedicated to the utilization of RGB sensors for tracking, also taking into account interactions of hands with external objects [39], [40], [41]. Many times schemes were developed for a non-specific number of cameras [38], [39], [21], [41]. 2-D image features as skin color, edges but also 3-D visual hulls are utilized for foreground detection. The hand models used are usually composed of geometrical primitives, but also Sum of Anisotropic Gaussians (SoAG) primitives have been used [30]. In [42], a synthetic image dataset is also created. The reported precision ranges from 3 to 20 mm.

From 2017 and after, tracking has been based on LMC sensors. Researchers focus on the calibration of the multisensory system, creating manual, semi-automatic [43] or self-calibration [46] methods. The orientation of the hand is frequently used for weighting the contribution of each sensor to a data-fusion result [2], [3], [43], while also comparisons of the efficiency of single versus multisensor structures are performed [45], [46], resulting always to the superiority of multiple sensor systems. In [47] data selection between the

TABLE VI: Key information for each described paper. The acronyms used at this table are, in alphabetical order: BLSTM-NN: Bidirectional Long-Short Term Memory Newral Network, CHMM: Coupled Hidden Markov Model, CNN: Convolutional Newral Network, CPSR: Corresponding Point Set Registration, CSV: C-Support Vector Classification, FLF: Feature Level Fusion, GA: Genetic Algorithm, G-N: Gauss-Newton GP: Geometrical Primitives models HCI: Human-Computer Interactions, HMM: Hidden Markov Model, ICP: Iteratice Closest Point, IK: Inverse Kinematics, ILO: Iterative Local Optimization, ISA: Interactive Simulated Annealing, KF: Kalman Filter, LBS: Linear Blend Skinning model, LDA: Linear Discriminant Analysis, LSF: Least Square Fitting, LSTM-RNN: Long Short-Term Memory Recurrent Newral Network, M: Medical purpose, ND: not described, NN: Nearest Neighbour, PCA: Principal Component Analysis, PPN: Perimeter Peak Number, PSO: Particle Swarm Optimization,, RANSAC: RANdom SAmple Consonsus, RF: Random Forests, RLF: Raw Level Fusion, RM: Robot Manipulation, RMSD: Root-Mean Square Deviation, RVMs: Relevance Vector Machines, SA: Simulated Annealing, SDLF: Decision Level Fusion, SLR: Signal Language Recognition, SMM: Skinned Mesh model, SoAG: Sum of Anisotropic Gaussians, SoG: Sum of Gaussians, SVD: Singular Value Decomposition, SVM: Support Vector Machines, VR: Virtual Reality, WTA: Winner-Takes-All.

		Ref.	Authors	Year	Sensors	Calibration	Optimization	Specific Classifiers	Fusion	Model	Purpose
Hand Tracking	Homogeneous	[35]	Nirei et al.	1996	2 RGB	ND	SA	-	-	GP	ND
		[36]	Utsumi et al.	1999	5 RGB	ND	KF	-	DLF	-	VR
		[37]	Placidi et al.	2007	3 RGB	Rigid Transf.	-	-	RLF	SMM	M
		[38], [39]	Oikonomidis et al.	2010/11	RGB (ND)	ND	PSO	-	FLF	GP	ND
		[40]	Ballan et al.	2012	8 RGB	ND	Local Opt.	-	FLF	LBS	ND
		[21]	Oikonomidis et al.	2013	RGB (ND)	ND	PSO	-	FLF	GP	ND
		[41]	Wang et al.	2013	RGB (ND)	ICP, ISA	ISA	-	FLF	SMM	HCI
		[30]	Sridhar et al.	2014	5 RGB	ND	ILO	-	RLF	3D SoAG	ND
		[42]	Panteleris et al.	2017	1 RGB-Stereo Pair	ND	PSO	-	FLF	SMM	ND
		[2], [3]	Placidi et al.	2017/18	2 IR	Rigid Transf.	-	-	DLF	LMC	M
		[43]	Novacek et al.	2021	IR (ND)	Kabsh [44]	-	-	RLF	LMC	HCI
		[45]	Houston et al.	2021	3 IR	Kabsh [44]	KF	-	RLF	LMC	ND
		[46]	Ovur et al.	2021	2 IR	Self-Calibr.	KF	-	RLF	-	M
		[47]	Placidi et al.	2021	2 IR	SVD	-	-	RLF	LMC	M
	Heterogeneous	[29]	Sridhar et al.	2013	5 RGB & 1 RGB-D	RANSAC	SVM	-	DLF	3D SoG	ND
		[4]	Penelle et al.	2014	1IR & 1 RGB-D	-	CPSR	-	FLF	-	M
		[28]	Tzionas et al.	2016	8 RGB & 1 RGB-D	-	G-N Method	-	FLF	LBS	ND
		[48]	Bo et al.	2018	1 IR & 1 RGB-D	Zhang [49]	-	-	RLF	-	VR
		[50]	Zhang et al.	2019	1 IR & 1 RGB-D	ND	-	-	FLF	LMC	RM
		[51], [52]	Wu et al.	2019	1 IR & 4 RGB-D	ND	-	-	FLF	Skeletal	VR
		[53]	Li et al.	2019	1 IR & 4 RGB-D	-	KF	-	FLF	LMC	RM
Gesture Recognition	Homogeneous	[54]	Stenger et al.	2001	RGB (ND)	ND	-	Unscented KF	-	GP	ND
		[55]	Malik et al.	2003	2 RGB	Rigid Transf.	-	PPN	FLF	-	HCI
		[56]	de Campos et al.	2006	3 and 4 RGB	-	-	RVM	RLF	GP	ND
		[57]	Schlattman et al.	2007	3 RGB	ND	-	-	FLF	-	HCI
		[58]	Wang et al.	2011	2 RGB	-	NN & IK	-	DLF	-	HCI
		[59]	Mihail et al.	2012	2 RGB-D	ND	-	NN	FLF	-	M
		[60]	Fok et al.	2015	2 IR	RMSD	-	HMM	FLF	LMC	SLR
		[61]	Rossol et al.	2015	2 IR	-	-	SVM	FLF	LMC	ND
		[62]	Mohandes et al.	2015	2 IR	ND	-	LDA	FLF & DLF	LMC	SLR
		[63]	Jin et al.	2016	2 IR	Self-Calibr.	-	-	RLF	LMC	RM
		[7]	Sun et al.	2016	2 IR	-	-	SVM	FLF	LMC	M
		[64]	Simon et al.	2017	RGB (ND)	RANSAC	-	CNN	DLF	Skeletal	ND
		[11]	Shen et al.	2019	3 IR	ICP	-	PCA	FLF	LMC	VR
		[65]	Kiselev et al.	2019	3 IR	-	-	LR & CSV & XGBClassifier	FLF	LMC	ND
		[66]	Qi et al.	2021	2 IR	Self-Calibr.	KF	LSTM-RNN	RLF	-	M
		[67]	Worrallo et al.	2021	2 IR	ND	-	-	DLF	LMC	VR
		[68]	Wang et al.	2021	5 IR	LSF & SVD	-	ND	FLF	LMC	VR
	Heterogeneous	[69]	Erden et al.	2014	3 PIR & 1 RGB	-	-	WTA hash based alg.	DLF	-	ND
		[70]	Marin et al.	2014	1 IR & 1 RGB-D	-	-	SVM	FLF	LMC	SLR
		[71]	Sreejith et al.	2015	1 IR & 1 RGB-D	-	-	Adaboost & Kinect SDK & HMM	-	LMC	HCI
		[72]	Craig et al.	2016	1 IR & 1 RGB-D	-	-	-	FL	-	VR
		[33]	Marin et al.	2016	1 IR & 1 RGB-D	RANSAC	-	SVM & RF	FLF	LMC	SLR
		[32]	Kumar et al.	2017	1 IR & 1 RGB-D	-	-	HMM & BLSTM-NN	FLF & DLF	LMC	SLR
		[31]	Kumar et al.	2017	1 IR & 1 RGB-D	-	-	HMM & CHMM	FLF & DLF	LMC	SLR
		[34]	d'Eusanio et al.	2020	1 IR & 1 RGB-D	-	-	CNN	FLF & DLF	-	HCI
		[73]	Yu et al.	2021	1 IR & 1 RGB-D	ND	ND	-	-	LMC	RM

sensors is for the first time performed separately for each joint and based on velocity information.

Regarding heterogeneous systems, after 2010 RGB-D and LMC data are used in a complementary way, with the RGB-D tracking the whole hand's movements while the LMC provides data for delicate finger movements [4], or weighted average fusion results are calculated [48], [50]. In a series of works of 2019 [51], [52], RGB-D sensors provide full-body information and are only combined with an LMC sensor for the hand tracking. Also, each sensor can provide different kinds of information, as in [53] where position data from a LMC are blend with velocity data from a Kinect.

From Table VI it can be observed that the number of published papers has raised a lot after 2010. In particular, more than 50% of the manuscript published in the last 12 years, were produced in the last 5. This number has grown especially because of the opportunities offered by the use of multiple heterogeneous sensors which greatly increased effectiveness and efficiency (from a spatial accuracy of about 20 mm and a temporal resolution of about 10 fps of the first years, we are actually obtaining a precision of about 2 mm and a frame rate of about 40 fps). This has allowed for ever more accurate tracking systems to be used for delicate interactive applications, ranging from quantitative medicine, rehabilitation and precision surgery, to remotely driven rovers and robots in dangerous (such as nuclear plants) or distant (such as Mars explorations) environments. Again from Table VI, it can be observed that, in the last generation of hand tracking systems, the tendency is to use homogeneous multiple IR sensors which are precise, fast, and not affected by the environmental lighting conditions. Their internal software implements efficient 3D recognition strategies which allow to obtain directly a numerical model of the hand. The resulting model allow to perform subsequent high-level operations, for occlusion reduction or model fusion, thus increasing furtherly the systems' robustness and FOV. Another important aspect derivable from Table VI, is that tracking is always considered as an optimization problem, aiming at discovering the best-fitting position of a structured hand model with respect a cloud of key-point measurements. For this reason, classification is almost absent.

B.2) Gesture recognition

In 2011 and 2012, systems using two RGB [58], and two RGB-D [59] sensors, respectively, were developed, for use on telecommanding. Regarding homogeneous systems, on the next years studies were concentrated on IR sensors, following the same trend as hand tracking, with the exception of [73] where a Kinect and a LMC sensor are used complimentarily, with data from only one sensor utilized for each time instant, based on an automatic distance-based or a manual switch technique. Artificial hands have also been used for creating robust, repetitive movements, for discovering the best sensor positioning and for acquiring ground truth information [61], [7], [11]. In [65] also the optimal number of sensors is investigated. Following this path, many Sign Language Recognition (SLR) schemes were developed [62], [60], [66]. Weighted averages of the contribution for each sensor are calculated using the internal accuracy provided by the LMC [66], [61] by Principal Component Analysis (PCA) [11] or by other fusion methods. Lately, LMC devices have been implemented on head-mount displays for VR applications [67], [68].

Schemes using different kinds of sensors either merge the data or utilize each sensor separately. In the works of this survey, some gestures are recognized from the Kinect and some from the LMC, simultaneously, by two different classifiers in [71], while in [70] and [33] the classification algorithm is fed with different kinds of data from each device. Not only spatial but also velocity data from the different sensors can be fused [72]. Also, after mutual calibration, fused feature vectors can be fed to more than one classifier [33], [31], even for two-hand gestures [32]. By investigating diverse systems' combinations, the authors of [34] have concluded that merging depth and IR data gives the optimal results.

In a manner similar with hand tracking, gesture recognition benefited from the increased availability of diverse sensors, though its growth has been lower than tracking after the appearance of 3D sensors. This has occurred because the previous hand gesture systems, obtained with RGB sensors, already had acceptable accuracy and efficiency. The introduction of 3D sensors has made it possible to expand the range of applications, for example from sign-language recognition to automotive and remote control setups, and to increase efficiency (mostly with regard to temporal resolution), at lower costs. Similarly with hand tracking, and for the same reasons, gesture recognition has recently been moving towards the use of homogeneous IR sensors. In contrast, though, to hand tracking, for gesture recognition hand models are used little. Furthermore, in gesture recognition optimization is almost absent, while specific classifiers are used to compare measurements with previously labelled gestures and to reduce ambiguity.

IV. Conclusion

In this survey we provided a description of the devices and methods used in the last 12 years and some pillar works appeared before, concerning contactless multisensory hand monitoring architectures. We included a short description of the sensors, presented the common steps found in every method and we briefly reported the collected public datasets. Additionally, we organized the works according to the sensors they used (Table V) and to the problem they solved, tracking or gesture recognition, sorting them also by hardware choices, homogeneous or heterogeneous sensors, in ascending date (Table VI). The results (Table V) indicate that most of hand monitoring problems are solved with RGB or IR sensors and that (Table VI) RGB sensors were used approximately until 2015, when they have been supplanted by low-cost IR sensors. Furthermore, Table V indicates that RGB-D sensors are almost never used alone, but are very useful as supporting devices, especially in conjunction with IR sensors. Specifically, the role of RGB-D device have been the reduction of problems' complexity by using the depth information they provide to perform background removal and to enlarge the field of view of IR sensors for connecting hand-specific, IR-provided information with the RGB-D-provided wide-range information regarding the arm or the whole body.

For the future, the trend is to use IR sensors for both tracking and gesture recognition, though in different number for each modality. As shown in Table VI, hand tracking usually requires a larger number of sensors with respect to gesture recognition, having stringent demands on the correct spatial and temporal positioning of the hand model which is needed in order to reduce occlusions. Occlusions are, however, still an unresolved issue for contactless devices, since self-occlusions are very frequent in hand monitoring, but can be mitigated with the use of multiple sensors and hand model constraints.

It is worth pointing out that an excessive increase of the number of sensors is not proved beneficial, since it complicates the procedure and makes it more time consuming, without providing a real advantage in terms of occlusion reduction. More importantly, the multiple sensors could interfere and disturb each other. An experience-based rule could be, for hand tracking, to use no more than three sensors (two low-range IR sensors and one RGB-D supporting sensor to increase the FOV), while for gesture recognition the number should not exceed two, of the same type (IR). The supporting role of RGB-D sensors will be maintained until the development of precise long range IR sensors (increasing the range of IR sensors is already possible, at the price of accuracy degradation).

A final remark should me made to the recent methodological improvements provided by machine learning and deep learning approaches: they have been proved to be effective for gesture recognition, providing us with the potential for high accuracy and real-time recognition rates. Impressive results and new techniques are to be expected in this field, in the near future. However, new structured datasets for hand movements and poses should be presented to reduce biases related to the data collection process and fairness due related to the acquisition protocols.

ACKNOWLEDGMENT

The authors would like to thank the Italian Ministry of University and Research (Dottorato di Ricerca innovativo a caratterizzazione industriale n.2, PON 2014-2020) for supporting this work.

REFERENCES

[1] J. M. Rehg and T. Kanade, "Visual tracking of high dof articulated structures: an application to human hand tracking," in *European conference on computer vision*, pp. 35–46, Springer, 1994.

[2] G. Placidi, L. Cinque, A. Petracca, M. Polsinelli, and M. Spezialetti, "A virtual glove system for the hand rehabilitation based on two orthogonal leap motion controllers.," in *ICPRAM*, pp. 184–192, 2017.

[3] G. Placidi, L. Cinque, M. Polsinelli, and M. Spezialetti, "Measurements by a leap-based virtual glove for the hand rehabilitation," *Sensors*, vol. 18, no. 3, p. 834, 2018.

[4] B. Penelle and O. Debeir, "Multi-sensor data fusion for hand tracking using kinect and leap motion," in *Proceedings of the 2014 Virtual Reality International Conference*, pp. 1–7, 2014.

[5] E. Tarakci, N. Arman, D. Tarakci, and O. Kasapcopur, "Leap motion controller–based training for upper extremity rehabilitation in children and adolescents with physical disabilities: A randomized controlled trial," *Journal of Hand Therapy*, vol. 33, no. 2, pp. 220–228, 2020.

[6] Q. Wang, P. Markopoulos, B. Yu, W. Chen, and A. Timmermans, "Interactive wearable systems for upper body rehabilitation: a systematic review," *Journal of neuroengineering and rehabilitation*, vol. 14, no. 1, pp. 1–21, 2017.

[7] X. Sun, I. Cheng, and A. Basu, "Spatio-temporally optimized multi-sensor motion fusion," in *2016 IEEE International Symposium on Multimedia (ISM)*, pp. 425–430, IEEE, 2016.

[8] T. Hu, X. Zhu, X. Wang, T. Wang, J. Li, and W. Qian, "Human stochastic closed-loop behavior for master-slave teleoperation using multi-leap-motion sensor," *Science China Technological Sciences*, vol. 60, no. 3, pp. 374–384, 2017.

[9] C. Keskin, F. Kirac, Y. E. Kara, and L. Akarun, "Randomized decision forests for static and dynamic hand shape classification," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 31–36, IEEE, 2012.

[10] P. Kumar, R. Saini, P. P. Roy, and U. Pal, "A lexicon-free approach for 3d handwriting recognition using classifier combination," *Pattern Recognition Letters*, vol. 103, pp. 1–7, 2018.

[11] H. Shen, X. Yang, H. Hu, Q. Mou, and Y. Lou, "Hand trajectory extraction of human assembly based on multi-leap motions," in *2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 193–198, IEEE, 2019.

[12] L. S. Scimmi, M. Melchiorre, S. Mauro, and S. Pastorelli, "Experimental real-time setup for vision driven hand-over with a collaborative robot," in *2019 International Conference on Control, Automation and Diagnosis (ICCAD)*, pp. 1–5, IEEE, 2019.

[13] Y. Liang, G. Du, F. Li, and P. Zhang, "Markerless human-manipulator interface with vibration feedback using multi-sensors," in *2019 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pp. 935–940, IEEE, 2019.

[14] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 52–73, 2007.

[15] W. Jänig and R. Baron, "Complex regional pain syndrome: mystery explained?," *The Lancet Neurology*, vol. 2, no. 11, pp. 687–697, 2003.

[16] L. Dipietro, A. M. Sabatini, and P. Dario, "A survey of glove-based systems and their applications," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 4, pp. 461–482, 2008.

[17] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1995–2006, 2013.

[18] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 677–695, 1997.

[19] D. J. Sturman and D. Zeltzer, "A survey of glove-based input," *IEEE Computer graphics and Applications*, vol. 14, no. 1, pp. 30–39, 1994.

[20] S. Khamis, J. Taylor, J. Shotton, C. Keskin, S. Izadi, and A. Fitzgibbon, "Learning an efficient model of hand shape variation from depth images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2540–2548, 2015.

[21] I. Oikonomidis, N. Kyriazis, K. Tzevanidis, and A. A. Argyros, "Tracking hand articulations: relying on 3d visual hulls versus relying on multiple 2d cues," in *2013 International Symposium on Ubiquitous Virtual Reality*, pp. 7–10, IEEE, 2013.

[22] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[24] B. Kamiński, M. Jakubczyk, and P. Szufel, "A framework for sensitivity analysis of decision trees," *Central European journal of operations research*, vol. 26, no. 1, pp. 135–159, 2018.

[25] J. Kennedy, "Swarm intelligence," in *Handbook of nature-inspired and innovative computing*, pp. 187–219, Springer, 2006.

[26] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[27] A. D. Wilson, S. Izadi, O. Hilliges, A. Garcia-Mendoza, and D. Kirk, "Bringing physics to the surface," in *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pp. 67–76, 2008.

[28] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall, "Capturing hands in action using discriminative salient points and physics simulation," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 172–193, 2016.

[29] S. Sridhar, A. Oulasvirta, and C. Theobalt, "Interactive markerless articulated hand motion tracking using rgb and depth data," in *Proceedings of the IEEE international conference on computer vision*, pp. 2456–2463, 2013.

[30] S. Sridhar, H. Rhodin, H.-P. Seidel, A. Oulasvirta, and C. Theobalt, "Real-time hand tracking using a sum of anisotropic gaussians model," in *2014 2nd International Conference on 3D Vision*, vol. 1, pp. 319–326, IEEE, 2014.

[31] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "Coupled hmm-based multi-sensor data fusion for sign language recognition," *Pattern Recognition Letters*, vol. 86, pp. 1–8, 2017.

[32] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing*, vol. 259, pp. 21–38, 2017.

[33] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with jointly calibrated leap motion and depth sensor," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 14991–15015, 2016.

[34] A. D'Eusanio, A. Simoni, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara, "Multimodal hand gesture classification for the human–car interaction," in *Informatics*, vol. 7, p. 31, Multidisciplinary Digital Publishing Institute, 2020.

[35] K. Nirei, H. Saito, M. Mochimaru, and S. Ozawa, "Human hand tracking from binocular image sequences," in *Proceedings of the 1996 IEEE IECON. 22nd International Conference on Industrial Electronics, Control, and Instrumentation*, vol. 1, pp. 297–302, IEEE, 1996.

[36] A. Utsumi and J. Ohya, "Multiple-hand-gesture tracking using multiple cameras," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 1, pp. 473–478, IEEE, 1999.

[37] G. Placidi, "A smart virtual glove for the hand telerehabilitation," *Computers in Biology and Medicine*, vol. 37, no. 8, pp. 1100–1107, 2007.

[38] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Markerless and efficient 26-dof hand pose recovery," in *Asian Conference on Computer Vision*, pp. 744–757, Springer, 2010.

[39] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints," in *2011 International Conference on Computer Vision*, pp. 2088–2095, IEEE, 2011.

[40] L. Ballan, A. Taneja, J. Gall, L. Van Gool, and M. Pollefeys, "Motion capture of hands in action using discriminative salient points," in *European Conference on Computer Vision*, pp. 640–653, Springer, 2012.

[41] Y. Wang, J. Min, J. Zhang, Y. Liu, F. Xu, Q. Dai, and J. Chai, "Video-based hand manipulation capture through composite motion control," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, pp. 1–14, 2013.

[42] P. Panteleris and A. Argyros, "Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 575–584, 2017.

[43] T. Novacek, C. Marty, and M. Jirina, "Project multileap: Fusing data from multiple leap motion sensors," in *2021 IEEE 7th International Conference on Virtual Reality (ICVR)*, pp. 19–25, IEEE, 2021.

[44] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 32, no. 5, pp. 922–923, 1976.

[45] A. Houston, V. Walters, T. Corbett, and R. Coppack, "Evaluation of a multi-sensor leap motion setup for biomechanical motion capture of the hand," *Journal of Biomechanics*, p. 110713, 2021.

[46] S. E. Ovur, H. Su, W. Qi, E. De Momi, and G. Ferrigno, "Novel adaptive sensor fusion methodology for hand pose estimation with multileap motion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–8, 2021.

[47] G. Placidi, D. Avola, L. Cinque, M. Polsinelli, E. Theodoridou, and J. M. R. Tavares, "Data integration by two-sensors in a leap-based virtual glove for human-system interaction," *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 18263–18277, 2021.

[48] L. Bo, C. Zhang, H. Cheng, and B. Baoxing, "Fingertip data fusion of kinect v2 and leap motion in unity," *Ingenierie des Systemes d'Information*, vol. 23, no. 6, p. 143, 2018.

[49] Z. Zhang, "Camera calibration with one-dimensional objects," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 7, pp. 892–899, 2004.

[50] B. Zhang, G. Du, W. Shen, and F. Li, "Gesture-based human-robot interface for dual-robot with hybrid sensors," *Industrial Robot: the international journal of robotics research and application*, 2019.

[51] Y. Wu, Y. Wang, S. Jung, S. Hoermann, and R. W. Lindeman, "Exploring the use of a robust depth-sensor-based avatar control system and its effects on communication behaviors," in *25th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–9, 2019.

[52] Y. Wu, Y. Wang, S. Jung, S. Hoermann, and R. W. Lindeman, "Towards an articulated avatar in vr: Improving body and hand tracking using only depth cameras," *Entertainment Computing*, vol. 31, p. 100303, 2019.

[53] C. Li, A. Fahmy, and J. Sienz, "An augmented reality based human-robot interaction interface using kalman filter sensor fusion," *Sensors*, vol. 19, no. 20, p. 4586, 2019.

[54] B. Stenger, P. R. Mendonça, and R. Cipolla, "Model-based 3d tracking of an articulated hand," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 2, pp. II–II, IEEE, 2001.

[55] S. Malik, "Real-time hand tracking and finger tracking for interaction csc2503f project report," *Department of Computer Science, University of Toronto, Tech. Rep*, 2003.

[56] T. E. de Campos and D. W. Murray, "Regression-based hand pose estimation from multiple cameras," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, pp. 782–789, IEEE, 2006.

[57] M. Schlattman and R. Klein, "Simultaneous 4 gestures 6 dof real-time two-hand tracking without any markers," in *Proceedings of the 2007 ACM symposium on Virtual reality software and technology*, pp. 39–42, 2007.

[58] R. Wang, S. Paris, and J. Popović, "6d hands: markerless hand-tracking for computer aided design," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 549–558, 2011.

[59] R. P. Mihail, N. Jacobs, and J. Goldsmith, "Static hand gesture recognition with 2 kinect sensors," in *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, p. 1, Citeseer, 2012.

[60] K.-Y. Fok, N. Ganganath, C.-T. Cheng, and K. T. Chi, "A real-time asl recognition system using leap motion sensors," in *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pp. 411–414, IEEE, 2015.

[61] N. Rossol, I. Cheng, and A. Basu, "A multisensor technique for gesture recognition through intelligent skeletal pose analysis," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 3, pp. 350–359, 2015.

[62] M. Mohandes, S. Aliyu, and M. Deriche, "Prototype arabic sign language recognition using multi-sensor data fusion of two leap motion controllers," in *2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15)*, pp. 1–6, IEEE, 2015.

[63] H. Jin, Q. Chen, Z. Chen, Y. Hu, and J. Zhang, "Multi-leapmotion sensor based demonstration for robotic refine tabletop object manipulation task," *CAAI Transactions on Intelligence Technology*, vol. 1, no. 1, pp. 104–113, 2016.

[64] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1145–1153, 2017.

[65] V. Kiselev, M. Khlamov, and K. Chuvilin, "Hand gesture recognition with multiple leap motion devices," in *2019 24th Conference of Open Innovations Association (FRUCT)*, pp. 163–169, IEEE, 2019.

[66] W. Qi, S. E. Ovur, Z. Li, A. Marzullo, and R. Song, "Multi-sensor guided hand gestures recognition for teleoperated robot using recurrent neural network," *IEEE Robotics and Automation Letters*, 2021.

[67] A. G. Worrallo and T. Hartley, "Robust optical based hand interaction for virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[68] Y. Wang, Y. Wu, S. Jung, S. Hoermann, S. Yao, and R. W. Lindeman, "Enlarging the usable hand tracking area by using multiple leap motion controllers in vr," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17947–17961, 2021.

[69] F. Erden and A. E. Cetin, "Hand gesture based remote control system using infrared sensors and a camera," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 4, pp. 675–680, 2014.

[70] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," in *2014 IEEE International conference on image processing (ICIP)*, pp. 1565–1569, IEEE, 2014.

[71] M. Sreejith, S. Rakesh, S. Gupta, S. Biswas, and P. P. Das, "Real-time hands-free immersive image navigation system using microsoft kinect 2.0 and leap motion controller," in *2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pp. 1–4, IEEE, 2015.

[72] A. Craig and S. Krishnan, "Fusion of leap motion and kinect sensors for improved field of view and accuracy for vr applications," *Virtual Reality Course Report*, 2016.

[73] J. Yu, M. Li, X. Zhang, T. Zhang, and X. Zhou, "A multi-sensor gesture interaction system for human-robot cooperation," in *2021 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, vol. 1, pp. 1–6, IEEE, 2021.
[74] G. Placidi, D. Avola, D. Iacoviello, and L. Cinque, "Overall design and implementation of the virtual glove," *Computers in Biology and Medicine*, vol. 43, no. 11, pp. 1927–1940, 2013.