This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469551, IEEE Transactions on Circuits and Systems for Video Technology

A Survey on 3D Hand Gesture Recognition

Hong Cheng, *Senior Member, IEEE,* Lu Yang, *Member, IEEE,* and Zicheng Liu, *Fellow, IEEE*

***Abstract*—3D hand gesture recognition has attracted increasing research interests in computer vision, pattern recognition and human-computer interaction. The emerging depth sensors greatly inspired various hand gesture recognition approaches and applications, which were severely limited in the 2D domain with conventional cameras. This paper presents a survey of some recent works on hand gesture recognition using 3D depth sensors. We first review the commercial depth sensors and public datasets which are widely used in this field. Then, we review the state-of-the-art research for 3D hand gesture recognition in four aspects: 3D hand modeling, static hand gesture recognition, hand trajectory gesture recognition and continuous hand gesture recognition. While the emphasis is on 3D hand gesture recognition approaches, the related applications and typical systems are also briefly summarized for practitioners.**

***Index Terms*—Depth Sensor, Hand Gesture Recognition, Skeleton Detection and Tracking, Begin-End Gesture Detection, Dynamic Time Warping**

I. Introduction

Hand gestures are elementary movements of a person's hands, and are the atomic communication components representing the thoughts of a person [1]. Evolutionary anthropologists tell us the use of hand gestures has been used since the beginning of human history and are much older than speech [2]. Moreover, hand gestures are natural, ubiquitous and meaningful part of spoken language, and researchers have claimed that gesture and sound form a tightly integrated system during human cognition [3]. Inspired by human interaction mainly by vision and sound, the use of hand gestures is one of most powerful and efficient ways in Human-Computer Interaction (HCI) [4].

There are three basic types of sensors which are capable of sensing hand gestures: mount based sensors, multi-touch screen sensors and vision based sensors. In the first case, accelerometers or gyros are used to capture the movement of hands and fingers [5]. Multi-touch screen sensors [6, 7] are suitable for mobile devices but limit the distance between users and computers. On the other hand, there are some advantages of vision based sensors [8]–[10] since they can be less cumbersome and uncomfortable than the mounted sensors due to no physical contact with users. Vision based sensors also provide much larger working distance than multi-touch screen sensors. However, the computational complexity is quite high for conventional vision based hand detection and tracking [11]. To handle this challenge, colored markers or data gloves have been employed to simplify the vision tasks [12]. Although these wearable land markers circumvent the skin segmentation [13], they place additional burden on users and could feel unnatural enough to perform hand gestures, which is the fatal weakness for HCI applications [14].

During the last five years, the progress of commercial 3D depth sensing technologies [15, 16] have greatly promoted the research of hand gesture recognition [17, 18]. An efficient hand segmentation step is the primary step in hand gesture recognition approaches. 3D depth information can be used to extract hand silhouettes for robust hand gesture recognition in a comfortable and efficient way by simply thresholding a depth map to isolate the hands. The threshold can be approximated according to the depth of the face [17]. A survey of depth-based hand segmentation techniques can be found in [19].

Generally, the depth based hand gesture recognition approaches fall into three categories: static hand gesture recognition [20], hand trajectory gesture recognition [21, 22] and continuous hand gesture recognition [23, 24]. All of three kinds of hand gestures can leverage the 3D hand modeling [25, 26] for fine-grained gesture (*e.g.* finger movement) recognition. Static gestures can represent digits while trajectory gestures describe strokes and letters. The continuous hand gesture recognition can determine when a gesture starts and when it ends from hand motion trajectories. Hence, we can realize the online understanding of hand gestures.

Considered as a kind of human motion [27] in our daily life, hand gestures have been investigated by human motion analysis [28]–[30], which pay more attention to the full-body human poses and activities [31]. Recently, there are several surveys which emphasize on general hand gesture recognition [8, 10]. Different from above literatures giving broad and general gesture recognition reports, we mainly focus on 3D hand gesture recognition and thoroughly summarize the very recent progress in the static, trajectory and continuous viewpoints, respectively. This survey can cover most emerging 3D hand gesture recognition approaches and systems, and reflect the research trend towards the practical HCI applications from assistive robots to wearable devices with the First Person Vision (FPV) technology [32, 33].

The remainder of the paper is organized as follows. Sec. II gives a glance of popular depth sensors which are widely accepted as basic sensing equipments in the community. Sec. III summarizes various 3D hand gesture datasets as benchmarks for experimental evaluations. The articulated 3D hand modeling by Degrees of Freedom (DoF) are reviewed in Sec. IV. The three core topics: static hand gesture recognition, hand trajectory gesture recognition and continuous hand gesture recognition are presented in Sec. V, Sec. VI and Sec. VII, respectively. The 3D hand gesture based practical applications and systems are introduced in Sec. VIII. Finally, Sec. IX concludes the paper. The overview of the surveyed 3D hand gesture recognition techniques is shown

Hong Cheng and Lu Yang are with both Center for Robotics and School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. Zicheng Liu is with the Microsoft Research Redmond, One Microsoft Way, Redmond, WA 98052, USA.
The Corresponding Author is Hong Cheng (hcheng@uestc.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469551, IEEE Transactions on Circuits and Systems for Video Technology

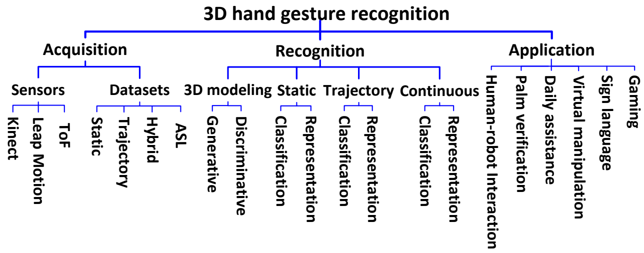


Fig. 1. The overview of 3D hand gesture recognition techniques

TABLE I
THE TAXONOMY OF SURVEYED TECHNIQUES WITH CORRESPONDING PUBLICATIONS

Publication	Sensor	Description
Acquisition		
Zhang (2012) [15]	Kinect	Introduction of Kinect
Nair *et al.* (2013) [16]	ToF	Survey of ToF sensors
Potter *et al.* (2013) [34]	Leap Motion	Leap Motion's use
Ren *et al.* (2011) [20]	Kinect	Static hand posture
Guyon *et al.* (2012) [35]	Kinect	Chalearn gesture challenge
Wang *et al.* (2012) [36]	Kinect	3D ASL dataset
Recognition		
Shotton *et al.* (2011) [37]	Kinect	Skeletal body parts
Keskin *et al.* (2013) [38]	Kinect	3D hand parts
de La Gorce *et al.* (2011) [39]	2D camera	3D hand modeling from 2D
Oikonomidis *et al.* (2010) [26]	Multi-camera	3D hand pose with 26 DoF
Oikonomidis *et al.* (2011) [40]	Kinect	Efficient 3D hand tracking
Hamer *et al.* (2009) [41]	Range camera	3D object manipulating
Oikonomidis *et al.* (2011) [42]	Multi-camera	3D hand with an object
Oikonomidis *et al.* (2012) [43]	Kinect	3D interacting hands
Ballan *et al.* (2012) [44]	Multi-camera	Interacting with an object
Van den Bergh *et al.* (2011) [45]	Kinect	Pointing gestures
Cheng *et al.* (2013) [4]	Kinect	Image-to-class DTW
Xu *et al.* (2012) [46]	Kinect	Robot navigation
Zhang *et al.* (2013) [47]	Kinect	3D daily hand gestures
Reyes *et al.* (2011) [48]	Kinect	Begin-end gesture
Cheng *et al.* (2014) [49]	Kinect	Continuous hand gestures
Application		
Kurakin *et al.* (2012) [18]	Kinect	Sign language recognition
Chai *et al.* (2009) [50]	Range camera	Virtual manipulation
Son *et al.* (2013) [51]	Kinect	Daily assistance
McKeague *et al.* (2013) [52]	Kinect	Human-robot interaction

in Fig. 1 with representative literatures in Table I. With the bottom-up taxonomy, all related techniques are classified into three categories: acquisition, recognition and application. The acquisition works involve various 3D sensors and datasets, which are the foundations of the following gesture recognition. The key components of recognition techniques are feature representation and classification. Finally, the recognized hand gestures can be widely used in different applications. In next sections, we will review each category in detail.

II. DEPTH SENSORS

Various 3D depth sensing technologies have been well reviewed by previous literatures [15, 16, 29], which mainly introduce the working mechanisms of different depth sensors[1]. In this survey, we focus on three popular depth sensors for hand gesture recognition. In next subsections, we will briefly review Kinect, Leap Motion and Time of Flight sensors (See Table II for brief comparisons). Although the conventional stereo camera such as Point Grey's Bumblebee can also sense the hand depth, its higher price restricts its applicability to the low-cost HCI application.

TABLE II
THE CONSUMER DEPTH SENSORS

Sensor	Resolution	Range	Accuracy	Description
Kinect 1.0	320×240	0.8-4.0m	4mm	20 body joints
Leap Motion	640×240	25-600mm	0.01mm	27 hand joints
Kinect 2.0 (ToF)	512×484	0.8-4.5m	1mm	25 body joints

A. Kinect

Since the release of Kinect 1.0 (for Xbox 360) in 2010, it has been the most popular low-cost depth sensor in the computer vision community. The associated OpenNI and SDK libraries enable the skeletal tracking of human body joints, which provide convenient information for gesture recognition [19]. Depending on whether the Kinect human skeleton [37] is used or not, the Kinect-based hand gesture recognition can be divided into two cases: skeleton based recognition and depth based recognition.

In the first case, the skeletal body joints especially the hand palm joints are utilized for fast hand detection and tracking [53]. Due to the advantage of fast and low-cost hand extraction, hand trajectory gesture recognition has been applied to various HCI applications such as TV control [54]. In addition to skeleton based hand trajectory gesture, the hybrid gestures [55, 56] containing both trajectory and static gestures are also recognized leveraging the Kinect skeleton system. With the power of Kinect skeleton tracking libraries, 3D hand motion related sign language vocabularies have been successfully recognized and translated for American [57] and Chinese [58].

In the second case, the Kinect sensor is considered as the pure depth sensing device and all sequential procedures are based on the RGB-D data without activating the inherent skeleton system [59]. Consequently, not only the hand gesture recognition can be applied to on-line applications, but also to the off-line recorded RGB-D videos. The independence of on-line skeleton library is crucial for developing and evaluating hand gesture recognition approaches on public RGB-D datasets. In contrast to the skeleton provided by the official libraries which only identify the palm center, finer hand modeling such as finger joints can be accurately achieved using the RGB-D frames [40, 60, 61]. We will further discuss 3D hand modeling in Sec. IV. Among the works in this case, the most straightforward use of RGB-D data is the depth-based hand [45, 62, 63]/finger [64, 65] detection and tracking [66, 67]. After the hand detection and tracking, either static hand gesture recognition [68]–[70] or hybrid hand gesture recognition [71, 72] can be applied.

B. Leap Motion

Unlike the Kinect sensor which captures the full body depth, the newly released Leap Motion focuses on the accurate 3D hand positioning. The Leap Motion sensor can detect hands and fingers with the accuracy around 0.01mm. After its release

[1]http://www.microsoft.com, https://www.leapmotion.com/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469551, IEEE Transactions on Circuits and Systems for Video Technology

in February of 2013, many researchers have considered it as a promising 3D sensor particularly suitable for 3D hand gesture recognition. There have been several pioneer works using Leap Motion for the HCI applications. Leap Motion sensor was firstly used to track the users' fingers to enable a hand-controlled interface in a virtual environment [73]. Meanwhile, the Leap motion was also employed in the 3D molecular graphic systems [74]. Although Leap motion has the potential to recognize more complex 3D hand gestures such as the Brazilian sign language [75], there is still few work in this direction. Since the accuracy [76] and the suitability of the Leap Motion controller for sign language recognition [34] have been well reviewed, we expect these evaluations for Leap Motion can help researchers to further develop its toolkits for the 3D hand gesture recognition application.

C. Time-of-Flight (ToF) Sensors

Before the release of Kinect and Leap Motion, the ToF camera (*e.g.* Zcam [77]) had been considered as a low-cost depth measurement device [78] and was widely used in the fields of computer vision and HCI. Particularly, the newly released Kinect 2.0 is also a ToF camera, which is different from Kinect 1.0 that uses light coding technique. ToF based 3D hand gesture recognition system can be used to control a robot by point gestures [79]. Regarding the static hand posture recognition, ToF sensors have been reported to be quite efficient and robust to the hand's orientation, size and cluttered backgrounds [80]. Due to its accuracy and robustness for the distance measurement task, the ToF sensor has been used to generate the 3D point cloud and helped in recognizing complex hand trajectory gestures such as Polish sign language [81].

III. Datasets

Among numerous public datasets[2][3][4] in vision field, both 2D [82, 83] and 3D [84] human action datasets have been provided for body gesture research. While those datasets collect very limited hand gesture such as hand waving, 3D hand gesture datasets have attracted more research interests for solving real-world problems [85]. A comprehensive survey for human gesture datasets can be found in [86]. In next subsections, we will focus on the emerging 3D hand gesture datasets and will categorize them from static, trajectory and hybrid viewpoints, respectively (Table III). Furthermore, the well-known American Sign Language (ASL) dataset [87] and its derived versions will also be reviewed.

A. Static Gesture Datasets

The static 3D gestures datasets usually capture the palm and finger postures in the RGB-D domain, which can represent basic symbols such as Arabic numerals. The 10-Gesture dataset [20] was the first which collected 3D static hand gestures with a Kinect sensor. The dataset was collected from 10 subjects who performed 10 different hand poses using the black belt on the gesturing hand's wrist. Hence, the 10-Gesture dataset has 100 gesture samples in total. Each of the sample consists of a color image and a depth image. The dataset was quite challenging since it was collected in uncontrolled environments with cluttered backgrounds. Besides, the subject's gesture varies in hand orientation, scale and articulation.

Other public static hand gesture datasets include the ASL Finger Spelling dataset [88] and the UESTC-ASL dataset [4]. The former contains alphabet signs recorded from 4 different persons, amounting to a total of 48,000 samples. The latter has a digit sign dataset with 100 samples, which is recorded from 10 subjects in different orientations, depths, and scales. Particularly, the UESTC-ASL dataset is more challenging than the 10-Gesture dataset since no wrist belt is used to help the hand segmentation, leading to more practical but difficult gesture recognition.

B. Trajectory Gesture Datasets

The most popular and well-known trajectory gesture dataset is the ChaLearn Gesture dataset [35], which has been used in the series of ChaLearn Gesture Challenge. It mainly focuses on hand and arm gestures for human-machine interaction. Some gestures have full-body presentations, but most of them only have upper body. Gestures are separated by returning to a resting position. The ChaLearn Gesture dataset has become the most important benchmark to evaluate 3D hand trajectory gesture recognition systems. Another 3D trajectory gesture datasets is the MSRC-12 Kinect gesture data set [89], which consists of sequences of human movements, represented as body-part locations, and the associated 12 gestures. It contains trajectories of 20 joints estimated using the Kinect Pose Estimation pipeline. The gestures are mainly represented by the two-hand movements. There are some other action/activity datasets such as MSR Action 3D, MSR Daily Activity Dataset [90], UESTC-DHG [49] and LIRIS human activity dataset. More or less, they all provide several easy hand trajectory gestures in the RGB-D domain.

C. Hybrid Gesture Datasets

Sometimes, static hand pose and dynamic hand moving will together represent the gesture. The 3D Iconic Gesture Dataset contains many pictorial gestures, which depict entities, objects or actions. The subjects perform iconic gestures to refer to entities through embodying their shapes. People can gesture the outline of an object either with static hand palm or dynamic trajectories using their imagination. The Sheffield Kinect Gesture (SKIG) Dataset [92] collects 10 common categories of 3D hand gestures, which are performed with three hand postures: fist, index and flat. Consequently, both hand tracking and shape recognition should be processed for final gesture recognition. Although the IDIAP Hand pose/gesture datasets[5] include both static hand posture datasets [94] and hand trajectory posture datasets [95], only 2D gestures are collected without the RGB-D information.

[2]http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm
[3]http://www.cvpapers.com/datasets.html
[4]http://datasets.visionbib.com/
[5]http://www.idiap.ch/resource/gestures/

TABLE III
SUMMARY OF 3D HAND GESTURE DATASETS

Dataset	Description	Availability
Static Gesture Datasets		
10-Gesture (2011) [20]	Digit gestures with wrist belt	http://eeeweba.ntu.edu.sg/computervision/people/home/renzhou/HandGesture.htm
ASL Finger Spelling (2011) [88]	Alphabet gestures	http://lifeprint.com/
UESTC-ASL (2013) [4]	Digit gestures	http://www.uestcrobot.net/?q=download
Trajectory Gesture Datasets		
ChaLearn (2012) [35]	Upper body gestures	https://www.kaggle.com/c/GestureChallenge
MSRC-12 (2012) [89]	Two-hand gestures	http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/
MSR Daily Activity (2012) [90]	Daily gestures	https://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm
UESTC-DHG (2013) [49]	Stroke gestures	http://www.uestcrobot.net/?q=download
LIRIS (2014) [91]	Annotated gestures	http://liris.cnrs.fr/voir/activities-dataset/
Hybrid Gesture Datasets		
3DIG	Iconic gestures	http://projects.ict.usc.edu/3dig/
SKIG (2009) [92]	Static and trajectory hand postures	http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm
ASL Datasets		
MSR Gesture3D (2012) [18, 36]	12 ASL gestures	http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm
3D Body Part Detection Video [93]	Both one-handed and two-handed signs	Not available

D. American Sign Language (ASL)

American Sign Language (ASL) has been considered as "the fourth most-used language in the United States" since the early 1970s [96]. Although the 2D ASL Lexicon Video Dataset (ASLLVD) has collected about 1500 lexical signs [97] in the National Center for Sign Language and Gesture Resources[6], there is still very limited 3D ASL datasets publicly available for the research purpose. To our knowledge, there are only two public 3D ASL datasets. The first one is the MSR Gesture3D Dataset [18, 36], which includes 12 ASL trajectory gestures performed by 10 subjects. In this dataset, the hand part has been segmented as the hand detection result. The other dataset is the 3D Body Part Detection Video Dataset [93], which contains 1113 signs both one-handed and two-handed with the skeleton information generated by the Kinect sensor. The dataset is expected to include most of the 3,000 signs found in the Gallaudet Dictionary of American Sign Language and serve as a benchmark for assessing future 3D hand gesture recognition algorithms.

IV. 3D HAND MODELING

3D hand modeling and tracking estimate the articulated hand poses and motions [98]. These are key technologies in many HCI applications such as robot surgery, virtual keyboard and sign language recognition [99]. According to the taxonomy proposed by [32], hand pose estimation approaches can be classified in two categories: discriminative approaches and generative approaches. Moreover, the hybrid 3D modeling which combines discriminative modeling and generative modeling has been also proposed. Next, we will review various 3D modeling approaches in the context of 3D hand pose estimation and tracking with respect to these three categories.

A. Discriminative 3D Hand Modeling

Discriminative approaches do not constrain the 3D hand model with explicit degrees of freedom. Instead, they train the classifiers which inversely map appearance-specific hand pixel features to unknown hand parameters (part label, pose parameter *etc.*). The classifiers are usually learned off-line from a large set of training samples. Besides, most of those approaches employ the decision tree based classifiers to accelerate the estimation in each frame independently.

Part Based Modeling. In the popular work of [37], Shotton *et al.* proposed the human pose recognition technique, which finally extracts skeletal body parts with a Kinect sensor. They constructed the randomized decision forests with 300k synthetic training images per tree to determine the body part label for each pixel in the depth domain. Inspired by this seminal work, Keskin *et al.* [38, 61] generalized the idea of [37] to the 3D hand pose estimation using depth data. They used a 3D skinned mesh model as the hand pose generator to synthesize up to 200k training images to learn the Randomized Decision Trees (RDT) for hand parts. Following [37], the simple depth comparison feature for each pixel x in the depth image I can be computed as:

$$f_{u,v}\left(I,x\right) = I\left(x + \frac{u}{I\left(x\right)}\right) - I\left(x + \frac{v}{I\left(x\right)}\right) \quad (1)$$

where the offsets u, v are normalized by the pixel depth to ensure the feature is depth invariant.

During the training step, the features of training samples are input to the RDTs to learn the optimum offsets at each leaf node. After the training of all trees, the posterior probability $P\left(c|I,x\right)$ of part label c with the depth pixel x can be inferred by tree nodes. The final classification score is given by averaging all the distributions together in each part's forest. The pixel will be labeled as the specific hand part which provides the highest score. To further extract the hand skeleton, mean-shift algorithm can be used to find the centroid of each part [37, 61]. However, the part forests are trained in a large synthetic dataset with considerable memory requirements.

Efficient Part Labeling. Based in the work of [61], Keskin *et al.* further proposed the Shape Classification Forest (SCF) to classify hand shapes rather than parts [100]. The pixel feature computation here is exactly same as Eqn. (1). Every pixel of the input depth image will vote for a single shape label for the hand image:

$$c^{*} = \text{argmax}_{c} \frac{1}{N} \sum_{p=1}^{N} P\left(c|I,x_{p}\right) \quad (2)$$

[6]http://www.bu.edu/asllrp/ncslgr.html

where N is the number of foreground hand pixels, c is the hand shape label.

With the power of shape recognition, the RDTs can be trained with limited amount of variation, leading to the more efficient and accurate hand part estimation. To get benefit from both synthetic and realistic data, Tang *et al.* [101] proposed the Semi-supervised Transductive Regression (STR) forest which associated the sparsely labeled target domain (realistic depth images) with a fully labeled source domain (synthetic depth image). By considering the realistic-synthetic discrepancies, the STR forest can improve the accuracy of hand pose estimation with a low-labeling cost. The depth feature defined in Eqn. (1) is simple but sensitive to background changes. In order to generate background independent pixels features, Yao *et al.* fused three features: shape feature, surface feature and position feature from both RGB and depth cameras [102], which will be fed into the random forest classifier for part labeling. In contrast to [37, 61], they employed the color glove to collect real labeled training data rather than synthetic data to circumvent the non-trivial sample rendering.

Although the discriminative approaches are based on single frame such that there exists no track drifting issue and can be performed in real-time, they require a large set of high-quality training data to train the hand part recognizers. The lack of kinematic constraint makes them less robust to articulate motions and self occlusions.

B. Generative 3D Hand Modeling

The generative approaches are popular among recent state-of-the-art 3D hand modeling and tracking works. They are also called model-based hand tracking approaches, which try to fit the explicit 3D DoF model to the observed hand data [103]. Note that either 2D or 3D hand images can be used to estimate the 3D hand model.

Hand Modeling from 2D Images. de La Gorce *et al.* proposed generative approaches [39, 104] to estimate 3D hand poses from monocular 2D images. In [104], they modeled the hand by an articulated kinematic tree with 28 DoF. The original model was rendered by ellipsoids and polyhedra to form the hand surface model. Then, the hand silhouette could be extracted by differentiating the parametrized hand surface model. Finally, the optimal hand pose was simultaneously estimated with the foreground-background segmentation task in a Maximum Likelihood framework by matching the synthesized silhouette with the observed frame. During the model fitting process, the kinematic constraints were also incorporated to prune the unrealistic parameter configurations and accelerate the optimization. The same authors further incorporated texture and shading information in the objective function to handle the ambiguities which can not be solely distinguished by the hand silhouette [39]. Compared to [104], the hand surface in [39] was more carefully synthesized by the triangulated mesh with computer graphics techniques such as shading model and texture mapping. Hence, the self-occlusion and time-varying illumination can be properly handled by the use of shading and texture. However, the 3D hand modeling from monocular 2D image requires rather high computational cost and limits its applications.

Single 3D Hand Modeling. Oikonomidis *et al.* proposed to recover the 3D hand pose by matching the 26 DoF hand model to multiple camera views [26]. The rendering of the 3D hand surface employed a sphere and a truncated cylinder as 3D basic shapes. Both hand silhouette and edge maps were computed as the observation cues to evaluate the consistency between the hypothesized pose and the observed multiple views by minimizing the objective function

$$E(h, M) = \sum_{I \in M} D(I, h, C(I)) + \lambda \cdot kc(h) \tag{3}$$

where h is the hand pose hypothesis, I is one of the image from the multi-view set M. D is the likelihood function which measures the dissimilarity between the projected hand surface (by camera matrix $C(I)$) and the observed image in the feature space. $kc(h)$ represents the kinematic constraint for the hand configuration.

In [26], the final optimization of Eqn. (3) was carried out with the Particle Swarm Optimization (PSO) technique and GPU implementation in real time. Their multicamera system consists of calibrated cameras with numbers raging from 2 to 8. Another multicamera based 3D hand modeling system was proposed by Ho *et al.* [105], in which different features were integrated to evaluate the observation likelihood function with particle filtering. In [40], Oikonomidis *et al.* proposed an efficient generative 3D hand tracking approaches using a Kinect sensor. The rendered 3D hand surface model is similar to the model of [26]. Moreover, the observation cues included both hand silhouette and depth information, leading to the more efficient objective function and simpler capturing system. With the power of Kinect sensor, the new objective function was based on the 3D structure and was robust to 2D illumination changes. In [106], Qian *et al.* modeled a hand simply using spheres. A fast cost function was proposed to measure the distance between the rendered model and the observed 3D point cloud. They further combined gradient based and stochastic optimization techniques to achieve the fast and robust hand tracking without using GPU.

3D Hand Modeling with a Manipulated Object. All the above 3D hand tracking approaches can provide promising hand pose estimation results when the hand is solely observed in isolation. However, they may not work well when the hand is manipulating an object. To handle the strong occlusion by the manipulated object, Hamer *et al.* proposed a system to recover the articulated 3D structure of the hand during object manipulation [41]. They employed the human hand model consisting of 27 bones and rendered each hand segment by a cylinder with a mesh approximating the skin. Both hand segmentation and depth information were used to evaluate the likelihood probability between the hand configuration and the observed RGB-D data. Moreover, the softened anatomical constraints and occlusion model were incorporated in a Markov Random Field (MRF), which was optimized by the Belief Propagation. They further modeled the object-specific prior from sparse training data to improve the robustness of their 3D hand tracker [107]. In [42], Oikonomidis *et al.* extended their work [26] by jointly estimating the 26 DoF hand pose and the model parameters of the manipulated object. The explicit parametric 3D modeling of the interacted object effectively represents the context of the hand and improves the pose estimation both for the hand and the object.

3D Interactive Hands Modeling. Maybe the most challenging scenario for 3D hand modeling task is the situation with interacting hands. Based on [40, 42], Oikonomidis *et al.* proposed a parametric model of the joint kinematics of two hands to track the full articulation of the strongly interacting hands [43]. The joint model has 54 DoFs (27 for each hand) with collision constraints and the RGB-D data is utilized with a Kinect sensor. They showed promising results of the optimized joint hand configuration. It was also demonstrated that the straightforward pose estimation for single hand tracking problem would result in a much lower accuracy in the interacting case. Ballan *et al.* [44] addressed the challenging problem of capturing the articulated motion of two hands that interact with each other and with an additional object. Regarding the observation cues, they employed multiple visual features such as edges, optical flow, salient points, and collisions to estimate the articulated pose within a single differentiable function. Compared to [43] which utilized Kinect sensor and PSO optimization, Ballan *et al.* adopted a multi-camera setup and the simple local optimization technique, which achieved lower pose estimation errors.

C. Hybrid 3D Hand Modeling

Recently, the hybrid 3D hand modeling frameworks which combine discriminative and generative approaches are also proposed. In [108], Xu *et al.* presented a three-step pipeline in which the first two steps were based on the discriminative depth features of [37] with the Hough forest regression model [109]. The initial hand pose estimation was followed by the final verification step which optimized a 27 DoF hand model in the generative way. At the same time, Sridhar *et al.* [110] proposed to combine a discriminative part-based pose retrieval approaches with a generative pose estimation approaches based on local optimization. Both works showed that the combination of discriminative and generative ideas can achieve state-of-the-art hand modeling accuracy as well as high efficiency.

V. Static Hand Gesture Recognition

After the 3D hand modeling, one may say it is straightforward to recognize the hand shape or posture due to the rich information in the estimated hand model. However, sometimes we are mainly interested in the meaning of the global hand appearance and the detailed hand skeletal model is not necessary. In this section, we will focus on 3D approaches which were developed for recognizing the static hand shapes.

A. Feature Representation

The hand features are usually extracted from the bounding box of the segmented hand. Both the RGB and depth information can help in detecting the hand area. Here we assume the Region of Interest (ROI) has been determined thus focus on the discriminative features for the ROI. Basically, the hand region's features could be roughly divided into three categories: low-level features, middle-level features, high-level features.

The low-level hand features can be generated either from the original 3D spatial domain or a transformed domain. Keskin *et al.* [111] formulated the hand feature by the set of pixel-wise depth features as defined in Eqn. (1) [37]. This depth feature accounts for different hand positions in a large feature space leading to the time-consuming matching. In order to significantly reduce the training time and the memory consumption, Kuznetsova *et al.* [112] proposed to use Ensemble of Shape Function (ESF) descriptor as the feature of the segmented hand region. The ESF descriptor consists of concatenated histograms which are generated with the random points in the point cloud. In addition to the features in the original depth domain, Haarlet coefficients [45], Gabor coefficients [88] and Flusser moment invariants [113, 114] have been extracted from transformed hand images to efficiently form rotation and intensity invariant hand features.

In contrast to low-level hand features which are based on the global hand information, the mid-level hand features are usually based on the local patch-level descriptors [115]. The hand region can be divided into rectangle cells [116] or cylinder sectors [117]. The distributions or histograms of points in all cells are concatenated to formulate the local descriptors. To leverage the strong shape information in the depth map, Zhang *et al.* [118] defined a 3D facet as a 3D local support surface associated with each 3D cloud point and proposed the Histogram of 3D Facets (H3DF) to represent the 3D hand shape. Not only local descriptors in regular grids, the local descriptor locates at the interest point has also been used to formulate the hand feature. Bagdanov *et al.* [119] represented the hand by the concatenation of the five SURF descriptors for a total of 640 dimensions.

Rather than the rough hand bounding box for the above low-level and middle-level features, the detailed hand contour or hand part is necessary for extracting the high-level hand features [120]–[122]. Ren *et al.* [20] represented the hand shape by the time-series curve where each finger corresponded to a segment of the curve. The time-series curve records the relative distance between each contour vertex to a center point and reserve the topological information as well. A similar time-series curve based hand feature was proposed by Cheng *et al.* in [4]. They generated fingerlets from those time-series curves by using different finger combinations. Each hand gesture was represented by the fingerlet ensemble. The high-level hand features also can be generated in a transformed domain. Liu *et al.* [123] explored the invariance from the hand contour and utilized Fourier descriptor, edge histogram and boundary moment invariants for the feature extraction. Besides, the number and the direction of the extracted finger tips have also been considered as features for the static hand gestures [124].

Note that the hybrid hand features can be generated by combining features in different levels [125]. The hybrid hand features can embed both local and global information of the hand region [126]. Liu *et al.* [127] adopted three types of geometric features which are translation, rotation and scale

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469551, IEEE Transactions on Circuits and Systems for Video Technology

invariant among fingers, palm and forearm. To evaluate different 3D hand features, Sorce *et al.* [128] compared hand mask with edges and concatenated SURF descriptors under different illuminations. Their experiments showed that edge is the worst unless some light-independent algorithm is used for the edge extraction. On the other hand, the SURF based feature is the best one in optimal lighting conditions. Those experimental results can help researchers to design more effective hand features by the fusion of different hand descriptors.

B. *Classifiers*

After the feature representation of the 3D hand, the proper classification is required to recognize the 3D hand posture. Among the static 3D hand gesture recognition works, there are four widely employed classification approaches: Support Vector Machines (SVM), Neural Network, Random Forests and Nearest Neighbor (NN) search (template matching).

SVM is the most popular classifier for the 3D static hand gesture recognition. It has been successfully applied to single-hand [117] and double-hand poses [127]. The LIBSVM module [129] can be easily used in the implementations. The kernel of a SVM can be either linear [118] or nonlinear [119]. The linear SVM [115] has been adopted to predict the class label with the ASL datasets [88]. In more complexed scenarios, the multi-class SVM [116] has successfully recognized 15 hand postures in the cluttered background with an average recognition rate of 95%. The SVM with linear kernel was also utilized with NTU Hand Digits dataset [20] and ASL Finger Spelling dataset. For nonlinear kernels, the SVM-RBF classifiers [114] were trained to recognize four different hand postures in a 3D medical touchless interface. We find that the SVM classifier is usually trained with the middle-level features.

Neural Network is still used for simple 3D static hand gesture recognition tasks. In [128], a neural network with back propagation detected the hand pose to recognize whether it is closed or not. An EWMA (Exponential Weighted Moving Average) noise reduction mechanism was used to suppress the noise effects of the neural network. Meanwhile, a neural network model was constructed to recognize four gesture stages [130]. In both works, the neural network only handled fundamental and limited hand gestures.

Random Forests (RF) [131] has also attracted attentions due to its fast training and competitive performance for handling large datasets and feature space. Furthermore, random forests intrinsically support multi-class classification thus can be easily parallelized. Hence, random forests based recognition can properly handle high-dimensional low-level and middle-level hand features. The multi-class random forests have been widely applied in practical systems for recognizing ASL Fingerspelling letters [88, 111, 112]. The random decision forest has been experimentally compared with the SVM classification with various 3D hand features [126]. Generally, the performance of random forests is dependent on the depth of the tree. The trade-off of accuracy and speed always needs to be properly considered in its implementation.

Nearest Neighbor Search or template matching can be used when the high-level discriminative 3D hand features are available. The key is the definition of the distance between the features. The Euclidean distance can be directly employed for the template matching of different hand gestures [123]. However, the generated hand shape features usually contain noises caused by local distortions, pose variations and inaccuracy depth maps. In order to robustly match the noisy hand features, Finger-Earth Mover's Distance (FEMD) was proposed [20] to measure the dissimilarity between time-curved hand features. To better handle the misalignment of hand features, Cheng *et al.* [4] proposed image-to-class Dynamic Time Warping (DTW) distance to distinguish the fingerlet features of 3D hand contours. We find that nearest neighbor search works quite well with the contour or boundary based hand features.

VI. Hand Trajectory Gesture Recognition

In contrast to the static hand gesture recognition which works on hand shapes, the hand trajectory gesture recognition considers the sequential data of hand trajectory and explores the temporal character of hand motion. In this section, we will survey the 3D hand trajectory gesture recognition approaches.

A. *Feature Representation*

The features used in hand trajectory gesture recognition can be divided into appearance based category and tracking based category.

The appearance based features rely on the local feature descriptors regardless of hand tracking or explicit motion trajectory. The features can be generated either in 2D color and depth frames or in the 3D temporal volumes. Wang *et al.* [132] concatenated body centroid, hand displacement and relative depth level of hand to represent the hand gesture in each depth frame. Similarly, Yang *et al.* [133] utilized groups of low-level image primitives such as region shape, proximity, or color to implicitly represent the hand as the salient object part of interest. The sequence of the frame-wise group primitives form the hand gesture feature in temporal domain without requiring a perfect hand segmentation. The early results of the ChaLearn Gesture Challenge [35] showed that all top ranking approaches were based on techniques making no explicit detection and tracking of humans or individual body parts. To handle the occlusion problem in depth maps, Wang *et al.* [36] proposed Random Occupancy Pattern (ROP) features which were extracted by sampling the depth space for the representation of hand action. As a special kind of human action, the hand trajectory gesture also can be represented by concatenating HOG based descriptor [134, 135]. All above features require no explicit hand tracking but discriminative appearance feature instead. The trajectory or motion of the hand can be implicitly embedded in those appearance cues.

The tracking based features require the tracking of the hand with hand's centroid position or the body skeleton. Hand positions, velocity, acceleration and chain code have been widely used to formulate trajectory features [136]. The hand position can be either determined by hand segmentation [137] or the body skeleton generated by NITE middleware with the Kinect sensor [138]. In addition to the palm position, the orientation or angle of the hand centroid in a 3D hand gesture

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469551, IEEE Transactions on Circuits and Systems for Video Technology

trajectory can also be used [46]. Miranda *et al.* [139] described the pose in each frame using a tailored angular representation of the skeleton joints. The key poses were identified with those descriptors. The gesture was represented as the sequence of key poses. For velocity based trajectory feature, Ren *et al.* [140] used the speed and direction of hand motion for the menu selection gesture. Similarly, Wu *et al.* [141] utilized the incremental changes of the three-dimensional coordinates in a unit time as the features. Those features rely on the valid hand tracking results and could fail when serious occlusion occurs.

More complex hand features have also been formulated based on the segmented and tracked hands. Zhang *et al.* [47] proposed a novel Edge Enhanced Depth Motion Map together with Histogram of Gradient descriptor to generate the vector representation of the hand trajectory gestures from the depth video. To ensure the generation of valid hand trajectory, Wang *et al.* [142] utilized the Potential Active Region (PAR). The hand gesture was represented by the segmented series of movements in the form of Motion History Images (MHI). Besides, more discriminative features were also proposed by projecting the motion trajectory to the higher-dimensional feature space [143]. Beh *et al.* [144] further composed the hand motion trajectory as a unique series of straight and curved segments. They proposed an automated process of segmenting gesture trajectories based on a simple set of threshold values in the angular change measure. The strokes and segments of hand trajectories played a crucial role in those approaches.

B. Classifiers

There are four main classifying approaches for recognizing hand trajectory gestures. Similar to the static hand gesture recognition, the hand trajectory gesture recognition can be carried out by the traditional classifiers such as SVM and NN matching. Considered as the sequential data, the hand trajectory features can also be handled by Hidden Markov Model (HMM) and DTW.

SVMs have been trained to classify the ASL trajectory gestures in the MSRGesture3D dataset [36, 47]. Five individual SVMs integrated by a second-stage linear SVM were utilized in the naturalistic CVRR-HANDS 3Ddataset [135]. For Nearest Neighbor search, a matching score between the model sequence and the input sequence was adopted as the measure to distinguish 39 different ASL signs and 7 hand actions in a two-view hand gesture dataset [133]. The maximum correlation coefficient was also used for the gesture recognition on Chalearn dataset [134]. For complex gesture features [143], the straightforward nearest neighbor (1-NN) with the Euclidean distance has also shown the powerful discriminative potential.

HMM has been widely used for the analysis of sequential data. It is very suitable to classify the tracking based gestures such as trajectory-based signed digits [138]. HMM was also applied to the robotic navigation [46] and sign language recognition [141, 144]. Compared to HMM, the main advantage of DTW is that it can automatically align the sequences which have different lengths and return the proper distance. Hence, DTW distance can be combined with k-NN classifiers for robust signed digits recognition [136, 137].

VII. Continuous Hand Gesture Recognition

For general hand trajectory gesture recognition, we usually assume the hand gesture video clip has been well segmented in the temporal domain. The focus is to recognize the single meaningful hand gesture in the video. However, the practical HCI applications require the continuous hand gesture recognition in video streams. This means that both the spatial segmentation and temporal segmentation are necessary to recognize the sequential hand gestures.

Following [23], we can roughly divide continuous hand gestures into direct approaches and indirect approaches depending on whether the explicit temporal segmentation proceeds or not. For 2D continuous hand gesture recognition, both direct [145] and indirect [23, 146, 147] approaches have been well investigated. Besides, continuous hand gesture recognition approaches which rely on the 3D accelerometer sensor [148] and touch screen [149] have also been proposed. Here, we will mainly emphasize on 3D continuous hand gesture recognition approaches with the depth data.

For direct 3D continuous hand gesture recognition, it is convenient to use low level motion features such as velocity and trajectory to detect abrupt changes for the spotting. Elmezain *et al.* [150] proposed a system to recognize continuous digit gestures in real-time using HMM. They first generated orientation dynamic features from spatio-temporal trajectories in depth domain and then quantized them to codewords. The segmentation of the continuous gestures were based on the detection of the zero-codeword, which actually detected the static velocity and the end point of the gesture. Kristensson *et al.* [151] presented a markerless gesture interface with a Kinect sensor. They defined the input zone for the gesture delimitation purpose. When the depth of the hand was below a set threshold, the hand was defined to be within the input zone. By this zoning technique, the beginning and the end of the gesture can be determined.

The main limitations of direct approaches are two-fold: 1) they require the pre-segmentation of gestures, which may delay the recognition result; 2) The requirement of begin/end signals (*e.g.* gesture interval) makes them not flexible in the HCI applications. Hence, the indirect approaches have received increasing interests due to its natural applicability and fast recognition for continuous hand gestures. Among the indirect continuous hand gesture recognition approaches, DTW is widely used for finding the matched gesture segment in the temporal domain. Lichtenauer *et al.* claimed that time warping and classification should be separated because of conflicting likelihood modeling demands [152]. They proposed to use Statistical DTW (SDTW) only for time warping, while classifying the warped with different statistical classifiers. The proposed hybrid approaches provided a significant improvement over HMM on the 3D Dutch Sign Language (DSL) dataset. To improve the recognition accuracy, Keskin *et al.* proposed a DTW based pre-clustering technique for 3D digit recognition using graphical models. To leverage the discriminativity of each hand part, Arici *et al.* [153] proposed a weighted DTW approaches that weighted joints by optimizing a discriminant ratio to improve the 3D hand-arm gesture recognition.

In the context of begin-end hand gesture recognition, Reyes *et al.* [48] presented a begin-end gesture recognition approaches using feature weighting in the DTW framework. The feature weighting approaches was proposed to improve the cost distance computation in the conventional begin-end DTW algorithm. They associated a discriminatory weight to each joint of the skeletal model depending on its participation in a particular gesture. DTW can also be applied to segment the gestures in the pre-processing step. Hernandez-Vela *et al.* [154] extended the BoW [155] model to Bag-of-Visual-and-Depth-Words (BoVDW) model for gesture recognition. The state-of-the-art RGB and depth features were fused to generate the final gesture feature, which was integrated in a continuous gesture recognition pipeline and the DTW algorithm was used to perform the begin-end segmentation of gestures. Besides, Bhuyan *et al.* [156] proposed to use a novel set of features with Conditional Random Fields (CRF) for gesture spotting to distinguish meaningful gestures from unintentional movements. Although their original approaches works for 2D continuous hand gesture recognition with skin-color based hand detection, it could be utilized for 3D scenario with hand tracking in the RGB-D space.

One of the most challenging practical problems for the continuous gesture recognition system is the subgesture problem when some gestures are similar to parts of other longer gestures. Subgesture problem is very common in real world gesture datasets (*e.g.* ASL and ChaLearn datasets). Recently, Cheng *et al.* proposed a Windowed Dynamic Time Warping (WDTW) approaches [49] for 3D continuous hand trajectory gesture recognition. In their work, a parameterized searching window was introduced in the cost matrix of traditional DTW approaches to detect the beginning and end of the specific gesture from an infinite trajectory gesture sequence. Hence, the continuous gesture recognition can be formulated into online parameter estimation of the searching window. Compared to the feature weighting DTW approaches in [154], the WDTW approaches can handle the subgesture problem by moving the searching window forward and judging whether this is a single gesture after some delay time.

VIII. Applications and Typical Systems

The 3D hand gestures recognition approaches are mainly used in five application domains: sign language recognition, virtual manipulation, daily assistance, palm verification, gaming and the emerging human-robot interaction.

Sign language recognition with hand gestures has been investigated for the ASL as mentioned in previous sections. For disable people who are deaf mute, a recognition system of sign language can greatly help them to keep in touch with others. Sun *et al.* [157] employed a novel two-Kinect system to collect the Japanese Sign Language (JSL) data. Two Kinect sensors were located perpendicularly to each other and the PCL Library was used to recognize the JSL gestures. Considering personal experiences, a fully automatic hand gesture recognition system can make people feel natural to use it. To approaches this goal, Kurakin *et al.* [18] proposed a real-time system for hand trajectory gesture recognition. The system was designed for practical ASL applications, fully automatic and robust to variations in speed and style as well as in hand orientations. This system is the first data-driven system that is capable of automatic hand gesture recognition. To further improve user experiences, Suau *et al.* [158] presented an online hand-based touchless interaction system named intAIRact. The intAIRact is user friendly and highly configurable with easy-to-learn hand gestures. Consequently, a user can trigger a large number of events by remembering 9 hand gestures and combining them with simple translations and rotations.

Virtual manipulation is the popular application of 3D hand gesture recognition due to its natural user interface for HCI tasks. Fraunhofer FIT developed the 3D Gesture-Based Interaction System[7] which is the noncontact gesture and finger recognition system. The system detects hand and finger positions in real-time and translates these into appropriate interaction commands. For picture gallery browsing, Chai *et al.* [50] leveraged accurate hand segmentation to greatly improve the interaction usability. To approaches user interfaces in dark environments, Lee *et al.* [159] presented an approaches for tracking hand rotation and various grasping gestures through an infrared camera. The 3D hand gesture recognition has also been combined with the 3D stereoscopic display to provide the immersive HCI experience in a virtual reality system with stereo vision [160]. In this system, an user can manipulate the virtual object in the 3D stereoscopy scene. The Kinect sensor was used to track the user's hand and render the virtual objects according to the user's viewpoint. Particularly, several recent systems involving Kinect-based touchless interaction with surgical images have been developed for operating-theatre practices [161]. The pioneer works include medical image manipulation systems at Sunnybrook Hospital in Toronto [162] and Guy's and St. Thomas' Hospital in London in London [161]. In these applications, 3D hand gestures vocabularies were carefully designed to realize the touchless image manipulation in sterile environments.

Daily assistance using hand gestures can help older people to perform Activities of Daily Living (ADLs) such as hand-washing [163]. The important senior gestures such as eating and drinking can also be monitored [164]. To make the smart home which facilitates aging in place, Yanik *et al.* [165] used Kinect depth data and growing neural gas algorithm for gesture based robot control. It was the initial effort towards the goal of assistive robot in which the response of the robot converged the user's desired response. For single-hand driving system, Son *et al.* [51] proposed a Kinect-based system that can help those people who have difficulties in moving one of their arms, to drive and control the vehicles with only one hand. It also has the potential to be applied to any wheeled vehicles. Regarding the typical systems, CMU's assistive robot 'HERB'[8] can help the user to accomplish the seamless teleoperation by predicting the user's manipulation intent [166]. Since older people also experience challenges in maintaining their home, care robots are considered as one option to support. Recently, Fischinger *et al.* [167] released a prototype care robot 'Hobbit' in which

[7] http://www.fit.fraunhofer.de/en/fb/cscw/projects/3d-multi-touch.html
[8] http://www.cmu.edu/herb-robot/

the 3D hand gesture interface was developed for the interaction of older people with the robot.

Palm verification is a key biometric technology for many security applications. The shape of the hand can be easily captured in a relatively user friendly manner by using 2D/3D cameras thus acceptable by the public. Amayeh *et al.* [168] proposed a component-based approaches to hand-based verification and identification. Their approaches utilized a 2D camera plus a lighting table to decomposite the hand silhouette into different regions corresponding to the back of the palm and the fingers. To identify the user's hand shape in free pose, Kanhangad *et al.* [169] proposed a contactless and pose invariant biometric identification system which utilized a 3-D digitizer to simultaneously acquire intensity and range images of the user's hand. Their approaches determined the orientation of the hand in 3D space and normalized the pose of the acquired RGB-D images to register the regions of interest for the matching based identification.

Gaming with interactive hand gestures have been significantly promoted by next-gen game consoles. Sony's Play Station and Nintendo's Wii are equipped with handy controllers (PS Move and Wiimote, respectively) for the player's hand tracking. Roccetti *et al.* [170] addressed the problem of differentiating the design of a gesture-based interface for a console from the problem of designing it for a public space setting. The most well-known somatic games involve Microsoft's Xbox with Kinect. There have been quite a few body-sensing games released in the consumer market.

Human-Robot Interaction (HRI) is the most important function for the emerging social robot which is able to interact with people using the natural gestures [171]. Hand gesture based interface offers a way to enable human to interact with robots more easily and efficiently [172]. Yin *et al.* [173] implemented a posture recognition system on a real humanoid service robot HARO-1 and experimental results demonstrated the effectiveness and robustness of the system. Among the set of gestures intuitively performed by humans, pointing gestures are especially important for interaction with robots. Nickel *et al.* [174] presented an approaches for recognizing pointing gestures in the context of human-robot interaction. The stereo camera was employed for the hand detection. The system aims at run-on gesture recognition in real-time and allow for the robot ego-motion. To handle the challenge of hand tracking in crowded and dynamic environments for HRI applications, McKeague *et al.* [52] proposed a sensor fusion based hand tracking algorithm for crowded environments. It significantly improved the accuracy of existing hand detectors, based on the RGB-D information. To support robust and efficient human-robot interaction for socially assistive robots, Michel *et al.* [175] proposed a 3D vision-based gesture recognition approaches which considered gestural vocabulary in the context of human-robot dialogue.

Particularly, to build an open and effective system for empowering natural modalities of HCI, Pedersoli *et al.* developed the first open source package for Kinect which targeted both static and trajectory based hand gestures in a unified framework [176]. The First Person Vision (FPV) technology [33] also has the potential to promote the emerging pervasive wearable devices (*e.g.* Google Glasses, Vuzix SmartGlass, Lenovo New Glass and Microsoft HoloLens) by hand gestures. Some pioneer works [177]–[179] in this direction has investigated hand detection and segmentation tasks with the ego-centric vision. Li and Kitani [179, 180] addressed the task of pixel-level hand detection with challenges of illumination changes and camera motion. The sequential classifier [181] and the superpixel classification [178, 182] techniques were also employed to explore the temporal and spatial coherence of hand gestures in first-person views. The real-time wearable hand gesture recognition [32] can be used for the HRI event detection (*e.g.* object manipulation) and the intention understanding. More hand gesture systems and applications have been well summarized in [183].

IX. Conclusions

In this paper we have given a comprehensive survey of the emerging progress on 3D hand gesture recognition. We discussed a variety of 3D hand gesture recognition aspects along with their applications. The survey reviewed important progress made in 3D depth sensors, datasets, 3D hand modeling, static hand gesture recognition, hand trajectory gesture recognition, continuous hand gesture recognition, related applications and typical systems. One of the major challenges is the online recognition of 3D hand gestures. The absence of explicit begin/end hints in practical scenarios will degrade the performance of traditional static and trajectory approaches. Hence, the continuous hand gesture recognition will attracted more attention due to its applicability. Other challenges include different meanings of similar hand gestures which need to be distinguished by fine-grained gesture recognition. In the future, we expect more tiny finger gestures can be well recognized leveraging more accurate depth sensors. Finally, we envision the booming of intelligent products using 3D hand gesture recognition for human-robot interaction purpose. Most of existed 3D hand gesture recognition works employ depth sensors with fixed position. However, users may move freely and disappear during the interaction with robots. We believe there will be much research room of interactive hand gesture recognition for human-robot interaction.

Acknowledgment

The authors would like to thank all the reviewers for their insightful comments. This work was partially supported by NSFC (No.61305033, 61273256), Fundamental Research Funds for the Central Universities (ZYGX2013J088, ZYGX2014Z009) and SRF for ROCS, SEM.

References

[1] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, 2011.

[2] R. Lefevre, *Rude Hand Gestures of the World: A Guide to Offending without Words*. San Francisco: Chronicle Books, 2011.

[3] S. D. Kelly, S. M. Manning, and S. Rodak, "Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education," *Language and Linguistics Compass*, vol. 2, no. 4, pp. 569–588, 2008.

[4] H. Cheng, Z. Dai, and Z. Liu, "Image-to-class dynamic time warping for 3D hand gesture recognition," in *IEEE ICME*, 2013.

[5] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang, "A framework for hand gesture recognition based on accelerometer and emg sensors," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 41, no. 6, pp. 1064–1076, 2011.

[6] D. Kammer, J. Wojdziak, M. Keck, R. Groh, and S. Taranko, "Towards a formalization of multi-touch gestures," in *ACM International Conference on Interactive Tabletops and Surfaces*, 2010.

[7] E. Hoggan, J. Williamson, A. Oulasvirta, M. Nacenta, P. O. Kristensson, and A. Lehtiö, "Multi-touch rotation gestures: Performance and ergonomics," in *ACM SIGCHI*, 2013.

[8] X. Zabulis, H. Baltzakis, and A. Argyros, "Vision-based hand gesture recognition for human-computer interaction," 2009, Chapter 34, The Universal Access Handbook.

[9] P. Buehler, M. Everingham, D. Huttenlocher, and A. Zisserman, "Upper body detection and tracking in extended signing sequences," *International Journal of Computer Vision*, vol. 95, no. 2, pp. 180–197, 2011.

[10] S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, pp. 1–54, 2012.

[11] X. Suau, J. Ruiz-Hidalgo, and J. Casas, "Real-time head and hand tracking based on 2.5D data," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 575–585, 2012.

[12] R. Y. Wang and J. Popovic, "Real-time hand-tracking with a color glove," in *ACM SIGGRAPH*, 2009.

[13] S. Phung, A. Bouzerdoum, and S. Chai, D., "Skin segmentation using color pixel classification: analysis and comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 148–154, 2005.

[14] R. Yang, S. Sarkar, and B. Loeding, "Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 462–477, 2010.

[15] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, 2012.

[16] R. Nair, K. Ruhl, F. Lenzen, S. Meister, H. Schäfer, C. Garbe, M. Eisemann, M. Magnor, and D. Kondermann, "A survey on time-of-flight stereo fusion," *Lecture Notes in Computer Science*, vol. 8200, pp. 105–127, 2013.

[17] M. Van den Bergh and L. Van Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," in *IEEE Workshop on Applications of Computer Vision*, 2011.

[18] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *European Signal Processing Conference*, 2012.

[19] J. Suarez and R. Murphy, "Hand gesture recognition with depth images: A review," in *IEEE RO-MAN*, 2012.

[20] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *ACM MM*, 2011.

[21] W.-G. Liou, C.-Y. Hsieh, and W.-Y. Lin, "Trajectory-based sign language recognition using discriminant analysis in higher-dimensional feature space," in *IEEE ICME*, 2011.

[22] C. Tran and M. Trivedi, "3-D posture and gesture recognition for interactivity in smart spaces," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 178–187, 2012.

[23] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1685–1699, 2009.

[24] S. Sarkar, B. Loeding, R. Yang, S. Nayak, and A. Parashar, "Segmentation-robust representations, matching, and modeling for sign language," in *IEEE CVPR Workshops*, 2011.

[25] J. Davis and M. Shah, "Toward 3-d gesture recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 13, pp. 381–393, 1999.

[26] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Markerless and efficient 26-DOF hand pose recovery," in *ACCV*, 2010.

[27] R. Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.

[28] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," *Lecture Notes in Computer Science*, vol. 8200, pp. 149–187, 2013.

[29] 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.

[30] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[31] Y. Song, J. Tang, F. Liu, and S. Yan, "Body surface context: A new robust feature for action recognition from depth videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 952–964, 2014.

[32] W. W. Mayol and D. W. Murray, "Wearable hand activity recognition for event summarization," in *IEEE International Symposium on Wearable Computers*, 2005.

[33] A. Betancourt, P. Morerio, C. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, In Press, 2015.

[34] L. E. Potter, J. Araullo, and L. Carter, "The Leap Motion controller: A view on sign language," in *Australian Computer-Human Interaction Conference*, 2013.

[35] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. Escalante, "Chalearn gesture challenge: Design and first results," in *IEEE CVPR Workshops*, 2012.

[36] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *ECCV*, 2012.

[37] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *IEEE CVPR*, 2011.

[38] C. Keskin, F. Kirac, Y. Kara, and L. Akarun, "Real time hand pose estimation using depth sensors," *Consumer Depth Cameras for Computer Vision*, pp. 119–137, 2013.

[39] M. de La Gorce, D. Fleet, and N. Paragios, "Model-based 3D hand pose estimation from monocular video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1793–1805, 2011.

[40] N. K. Iason Oikonomidis and A. Argyros, "Efficient model-based 3D tracking of hand articulations using Kinect," in *BMVC*, 2011.

[41] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool, "Tracking a hand manipulating an object," in *IEEE ICCV*, 2009.

[42] I. Oikonomidis, N. Kyriazis, and A. Argyros, "Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints," in *IEEE ICCV*, 2011.

[43] ——, "Tracking the articulated motion of two strongly interacting hands," in *IEEE CVPR*, 2012.

[44] L. Ballan, A. Taneja, J. Gall, L. Van Gool, and M. Pollefeys, "Motion capture of hands in action using discriminative salient points," in *ECCV*, 2012.

[45] M. Van den Bergh, D. Carton, R. de Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. Van Gool, and M. Buss, "Real-time 3D hand gesture interaction with a robot for understanding directions from humans," in *IEEE RO-MAN*, 2011.

[46] D. Xu, Y.-L. Chen, C. Lin, X. Kong, and X. Wu, "Real-time dynamic gesture recognition system based on depth perception for robot navigation," in *IEEE International Conference on Robotics and Biomimetics*, 2012.

[47] C. Zhang and Y. Tian, "Edge enhanced depth motion map for dynamic hand gesture recognition," in *IEEE CVPR Workshops*, 2013.

[48] M. Reyes, G. Dominguez, and S. Escalera, "Featureweighting in dynamic timewarping for gesture recognition in depth data," in *IEEE ICCV Workshops*, 2011.

[49] H. Cheng, J. Luo, and X. Chen, "A windowed dynamic time warping approach for 3D continuous hand gesture recognition," in *IEEE ICME*, 2014.

[50] X. Chai, Y. Fang, and K. Wang, "Robust hand gesture analysis and application in gallery browsing," in *IEEE ICME*, 2009.

[51] J. Son and A. Sowmya, "Single-handed driving system with Kinect," in *HCI*, 2013, pp. 631–639.

[52] S. McKeague, J. Liu, and G.-Z. Yang, "An asynchronous RGB-D sensor fusion framework using Monte-Carlo methods for hand tracking on a mobile robot in crowded environments," in *International Conference on Social Robotics*, 2013.

[53] K. Lai, J. Konrad, and P. Ishwar, "A gesture-driven computer interface using Kinect," in *IEEE Southwest Symposium on Image Analysis and Interpretation*, 2012.

[54] R.-D. Vatavu, "User-defined gestures for free-hand TV control," in *European Conference on Interactive TV and Video*, 2012.

[55] L. Gallo, A. Placitelli, and M. Ciampi, "Controller-free exploration of medical image data: Experiencing the Kinect," in *International Symposium on Computer-Based Medical Systems*, 2011.

[56] M. Caputo, K. Denker, B. Dums, and G. Umlauf, "3d hand gesture recognition based on sensor fusion of commodity hardware," in *Mensch & Computer*, 2012.

[57] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, “American sign language recognition with the Kinect,” in *International Conference on Multimodal Interfaces*, 2011.

[58] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, and M. Zhou, “Sign language recognition and translation with Kinect,” in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.

[59] M. Tang, “Recognizing hand gestures with Microsoft’s Kinect,” 2011, Department of Electrical Engineering, Stanford University.

[60] W. Zhao, J. Chai, and Y.-Q. Xu, “Combining marker-based Mocap and RGB-D camera for acquiring high-fidelity hand motion data,” in *ACM SIGGRAPH Eurographics Symposium on Computer Animation*, 2012.

[61] C. Keskin, F. Kirac, Y. Kara, and L. Akarun, “Real time hand pose estimation using depth sensors,” in *IEEE ICCV Workshops*, 2011.

[62] K. K. Biswas and S. Basu, “Gesture recognition using Microsoft Kinect,” in *International Conference on Automation, Robotics and Applications*, 2011.

[63] Y. Wen, C. Hu, G. Yu, and C. Wang, “A robust method of detecting hand gestures using depth sensors,” in *IEEE International Workshop on Haptic Audio Visual Environments and Games*, 2012.

[64] J. Raheja, A. Chaudhary, and K. Singal, “Tracking of fingertips and centers of palm using kinect,” in *International Conference on Computational Intelligence, Modelling and Simulation*, 2011.

[65] G. Du, P. Zhang, J. Mai, and Z. Li, “Markerless Kinect-based hand tracking for robot teleoperation,” *International Journal of Advanced Robotic Systems*, vol. 9, no. 36, pp. 1–10, 2012.

[66] V. Frati and D. Prattichizzo, “Using Kinect for hand tracking and rendering in wearable haptics,” in *IEEE World Haptics Conference*, 2011.

[67] U. Lee and J. Tanaka, “Hand controller: Image manipulation interface using fingertips and palm tracking with Kinect depth data,” in *Asia Pacific Conference on Computer Human Interaction*, 2012.

[68] Z. Ren, J. Meng, J. Yuan, and Z. Zhang, “Robust hand gesture recognition with Kinect sensor,” in *ACM MM*, 2011.

[69] Y. Li, “Hand gesture recognition using Kinect,” in *IEEE International Conference on Software Engineering and Service Science*, 2012.

[70] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, “Robust part-based hand gesture recognition using Kinect sensor,” *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

[71] M. Shiratuddin and K.-W. Wong, “Non-contact multi-hand gestures interaction techniques for architectural design in a virtual environment,” in *International Conference on Information Technology and Multimedia*, 2011.

[72] F. Pedersoli, N. Adami, S. Benini, and R. Leonardi, “XKin - EXtendable hand pose and gesture recognition library for Kinect,” in *ACM MM*, 2012.

[73] H. Regenbrecht, J. Collins, and S. Hoermann, “A Leap-supported, hybrid AR interface approach,” in *Australian Computer-Human Interaction Conference*, 2013.

[74] K. Sabir, C. Stolte, B. Tabor, and S. O’Donoghue, “The molecular control toolkit: Controlling 3D molecular graphics via gesture and voice,” in *IEEE Symposium on Biological Data Visualization*, 2013.

[75] A. Porfirio, K. Lais Wiggers, L. Oliveira, and D. Weingaertner, “LIBRAS sign language hand configuration recognition based on 3D meshes,” in *IEEE SMC*, 2013.

[76] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, “Analysis of the accuracy and robustness of the Leap Motion controller,” *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[77] G. J. Iddan and G. Yahav, “Three-dimensional imaging in the studio and elsewhere,” *Three-Dimensional Image Capture and Applications IV*, vol. 4298, no. 1, pp. 48–55, 2001.

[78] A. Kolb, E. Barth, and R. Koch, “ToF-sensors: New dimensions for realism and interactivity,” in *IEEE CVPR Workshops*, 2008.

[79] D. Droeschel, J. Stückler, and S. Behnke, “Learning to interpret pointing gestures with a time-of-flight camera,” in *International Conference on Human-robot Interaction*, 2011.

[80] S. Oprisescu, C. Rasche, and B. Su, “Automatic static hand gesture recognition using ToF cameras,” in *European Signal Processing Conference*, 2012.

[81] T. Kapuscinski, M. Oszust, and M. Wysocki, “Recognition of signed dynamic expressions observed by ToF camera,” in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications*, 2013.

[82] O. Kliper-Gross, T. Hassner, and L. Wolf, “The action similarity labeling challenge,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 615–621, 2012.

[83] T. Hassner, “A critical review of action recognition benchmarks,” in *IEEE CVPR Workshops*, 2013.

[84] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, “The i3DPost multi-view and 3D human action/interaction database,” in *Conference for Visual Media Production*, 2009.

[85] Y. Song, D. Demirdjian, and R. Davis, “Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database,” in *IEEE International Conference on Automatic Face and Gesture Recognition Workshops*, 2011.

[86] S. Ruffieux, D. Lalanne, E. Mugellini, and O. Abou Khaled, “A survey of datasets for human gesture recognition,” in *Human-Computer Interaction*, 2014.

[87] A. Martinez, R. Wilbur, R. Shay, and A. Kak, “Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language,” in *IEEE International Conference on Multimodal Interfaces*, 2002.

[88] N. Pugeault and R. Bowden, “Spelling it out: Real-time ASL fingerspelling recognition,” in *IEEE ICCV Workshops*, 2011.

[89] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin, “Instructing people for training gestural interactive systems,” in *ACM CHI*, 2012.

[90] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in *IEEE CVPR*, 2012.

[91] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia, and B. Sankur, “Evaluation of video activity localizations integrating quality and quantity measurements,” *Computer Vision and Image Understanding*, vol. 127, no. 0, pp. 14 – 30, 2014.

[92] L. Liu and L. Shao, “Learning discriminative representations from RGB-D video data,” in *International Joint Conference on Artificial Intelligence*, 2013.

[93] C. Conly, P. Doliotis, P. Jangyodsuk, R. Alonzo, and V. Athitsos, “Toward a 3D body part detection video dataset and hand tracking benchmark,” in *International Conference on Pervasive Technologies Related to Assistive Environments*, 2013.

[94] J. Triesch and C. von der Malsburg, “A system for person-independent hand posture recognition against complex backgrounds,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*,, vol. 23, no. 12, pp. 1449–1453, 2001.

[95] S. Marcel, O. Bernier, J. Viallet, and D. Collobert, “Hand gesture recognition using input-output hidden Markov models,” in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[96] R. E. Mitchell, T. A. Young, B. Bachleda, and M. A. Karchmer, “How many people use ASL in the United States? why estimates need updating,” *Sign Language Studies*, vol. 6, no. 3, pp. 306–335, 2006.

[97] A. Thangali, J. Nash, S. Sclaroff, and C. Neidle, “Exploiting phonological constraints for handshape inference in ASL video,” in *IEEE CVPR*, 2011.

[98] 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.

[99] D. Mohr and G. Zachmann, “A survey of vision-based markerless hand tracking approaches,” 2013, preprint submitted to Computer Vision and Image Understanding.

[100] C. Keskin, F. Kirac, Y. E. Kara, and L. Akarun, “Hand pose estimation and hand shape classification using multi-layered randomized decision forests,” in *ECCV*, 2012.

[101] D. Tang, T.-H. Yu, and T.-K. Kim, “Real-time articulated hand pose estimation using semi-supervised transductive regression forests,” in *IEEE ICCV*, 2013.

[102] Y. Yao and Y. Fu, “Real-time hand pose estimation from RGB-D sensor,” in *IEEE ICME*, 2012.

[103] Y. Wu, J. Lin, and T. Huang, “Capturing natural hand articulation,” in *IEEE ICCV*, 2001.

[104] M. de La Gorce and N. Paragios, “A variational approach to monocular hand-pose estimation,” *Computer Vision and Image Understanding*, vol. 114, no. 3, pp. 363–372, 2010.

[105] M.-F. Ho, C.-Y. Tseng, C.-C. Lien, and C.-L. Huang, “A multi-view vision-based hand motion capturing system,” *Pattern Recognition*, vol. 44, no. 2, pp. 443–453, 2011.

[106] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, “Realtime and robust hand tracking from depth,” in *IEEE CVPR*, 2014.

[107] H. Hamer, J. Gall, T. Weise, and L. Van Gool, “An object-dependent hand pose prior from sparse training data,” in *IEEE CVPR*, 2010.

[108] C. Xu and L. Cheng, “Efficient hand pose estimation from a single depth image,” in *IEEE ICCV*, 2013.

[109] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, “Efficient regression of general-activity human poses from depth images,” in *IEEE ICCV*, 2011.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469551, IEEE Transactions on Circuits and Systems for Video Technology

[110] S. Sridhar, A. Oulasvirta, and C. Theobalt, "Interactive markerless articulated hand motion tracking using RGB and depth data," in *IEEE ICCV*, 2013.

[111] C. Keskin, F. Kirac, Y. Kara, and L. Akarun, "Randomized decision forests for static and dynamic hand shape classification," in *IEEE CVPR Workshops*, 2012.

[112] A. Kuznetsova, L. Leal-Taixe, and B. Rosenhahn, "Real-time sign language recognition using a consumer depth camera," in *IEEE ICCV Workshops*, 2013.

[113] L. Gallo and A. Placitelli, "View-independent hand posture recognition from single depth images using PCA and Flusser moments," in *International Conference on Signal Image Technology and Internet Based Systems*, 2012.

[114] L. Gallo, "Hand shape classification using depth data for unconstrained 3D interaction," *J. Ambient Intell. Smart Environ.*, vol. 6, no. 1, pp. 93–105, Jan. 2014.

[115] X. Zhu and K. Wong, "Single-frame hand gesture recognition using color and depth kernel descriptors," in *ICPR*, 2012.

[116] H. Wang, Q. Wang, and X. Chen, "Hand posture recognition from disparity cost map," in *ACCV*, 2012.

[117] P. Suryanarayan, A. Subramanian, and D. Mandalapu, "Dynamic hand pose recognition using depth data," in *ICPR*, 2010.

[118] C. Zhang, X. Yang, and Y. Tian, "Histogram of 3D Facets: A characteristic descriptor for hand gesture recognition," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013.

[119] A. Bagdanov, A. Del Bimbo, L. Seidenari, and L. Usai, "Real-time hand status recognition from RGB-D imagery," in *ICPR*, 2012.

[120] M. Flasinski and S. Myslinski, "On the use of graph parsing for recognition of isolated hand postures of polish sign language," *Pattern Recognition*, vol. 43, no. 6, pp. 2249–2264, 2010.

[121] S. P. Priyal and P. K. Bora, "A robust static hand gesture recognition system using geometry based normalizations and Krawtchouk moments," *Pattern Recognition*, vol. 46, no. 8, pp. 2202–2219, 2013.

[122] Y.-T. Li and J. P. Wachs, "HEGM: A hierarchical elastic graph matching for hand gesture recognition," *Pattern Recognition*, vol. 47, no. 1, pp. 80–88, 2014.

[123] Y. Liu, Y. Yang, L. Wang, J. Xu, H. Qi, X. Zhao, P. Zhou, L. Zhang, B. Wan, D. Ming, D. Guo, B. Shi, and J. Wang, "Image processing and recognition of multiple static hand gestures for human-computer interaction," in *ICIG*, 2013.

[124] S. Qin, X. Zhu, H. Yu, S. Ge, Y. Yang, and Y. Jiang, "Real-time markerless hand gesture recognition with depth camera," in *PCM*, 2012.

[125] D. Uebersax, J. Gall, M. Van den Bergh, and L. Van Gool, "Real-time sign language letter and word recognition from depth data," in *IEEE ICCV Workshops*, 2011.

[126] D. Minnen and Z. Zafrulla, "Towards robust cross-user hand tracking and shape recognition," in *IEEE ICCV Workshops*, 2011.

[127] L. Liu, J. Xing, H. Ai, and X. Ruan, "Hand posture recognition using finger geometric feature," in *ICPR*, 2012.

[128] S. Sorce, V. Gentile, and A. Gentile, "Real-time hand pose recognition based on a neural network using microsoft kinect," in *IEEE International Conference on Broadband and Wireless Computing, Communication and Applications*, 2013.

[129] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[130] M. Asad and C. Abhayaratne, "Kinect depth stream pre-processing for hand gesture recognition," in *IEEE ICIP*, 2013.

[131] L. Breiman and E. Schapire, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[132] Y. Wang, T. Yu, L. Shi, and Z. Li, "Using human body gestures as inputs for gaming via depth analysis," in *IEEE ICME*, 2008.

[133] R. Yang and S. Sarkar, "Coupled grouping and matching for sign and gesture recognition," *Computer Vision and Image Understanding*, vol. 113, no. 6, pp. 663–681, 2009.

[134] B. Liang, "Gesture recognition using depth images," in *ACM ICMI*, 2013.

[135] E. Ohn-Bar and M. Trivedi, "The power is in your hands: 3D analysis of hand gestures in naturalistic video," in *IEEE CVPR Workshops*, 2013.

[136] M.-K. Sohn, S.-H. Lee, D.-J. Kim, B. Kim, and H. Kim, "A comparison of 3D hand gesture recognition using dynamic time warping," in *27th Conference on Image and Vision Computing New Zealand*, 2012.

[137] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard, and V. Athitsos, "Comparing gesture recognition accuracy using color and depth information," in *International Conference on PErvasive Technologies Related to Assistive Environments*, 2011.

[138] H.-M. Zhu and C.-M. Pun, "Real-time hand gesture recognition from depth image sequences," in *International Conference on Computer Graphics, Imaging and Visualization*, 2012.

[139] L. Miranda, T. Vieira, D. Martinez, T. Lewiner, A. Vieira, and M. Campos, "Real-time gesture recognition from depth data through key poses learning and decision forests," in *SIBGRAPI Conference on Graphics, Patterns and Images*, 2012.

[140] G. Ren and E. O'Neill, "3D selection with freehand gesture," *Computers & Graphics*, vol. 37, no. 3, pp. 10–120, 2013.

[141] Y.-K. Wu, H.-C. Wang, L.-C. Chang, and K.-C. Li, "Using HMMs and depth information for signer-independent sign language recognition," in *International Workshop, Multi-disciplinary Trends in Artificial Intelligence*, 2013.

[142] H. Wang, J. Fu, Y. Lu, X. Chen, and S. Li, "Depth sensor assisted real-time gesture recognition for interactive presentation," *Journal of Visual Communication and Image Representation*, vol. 24, no. 8, pp. 1458–1468, 2013.

[143] W.-Y. Lin and C.-Y. Hsieh, "Kernel-based representation for 2D/3D motion trajectory retrieval and classification," *Pattern Recognition*, vol. 46, no. 3, pp. 662–670, 2013.

[144] J. Beh, D. Han, and H. Ko, "Rule-based trajectory segmentation for modeling hand motion trajectory," *Pattern Recognition*, vol. 47, no. 4, pp. 1586–1601, 2014.

[145] H. Li and M. Greenspan, "Model-based segmentation and recognition of dynamic gestures in continuous video streams," *Pattern Recognition*, vol. 44, no. 8, pp. 1614–1628, 2011.

[146] L. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *IEEE CVPR*, 2007.

[147] H.-I. Suk, B.-K. Sin, and S.-W. Lee, "Hand gesture recognition based on dynamic Bayesian network framework," *Pattern Recognition*, vol. 43, no. 9, pp. 3059–3072, 2010.

[148] G. Raffa, J. Lee, L. Nachman, and J. Song, "Don't slow me down: Bringing energy efficiency to continuous gesture recognition," in *International Symposium on Wearable Computers*, 2010.

[149] P. O. Kristensson and L. C. Denby, "Continuous recognition and visualization of pen strokes and touch-screen gestures," in *Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling*, 2011.

[150] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, "A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory," in *ICPR*, 2008.

[151] P. O. Kristensson, T. Nicholson, and A. Quigley, "Continuous recognition of one-handed and two-handed gestures using 3D full-body motion tracking sensors," in *ACM International Conference on Intelligent User Interfaces*, 2012.

[152] J. Lichtenauer, E. Hendriks, and M. Reinders, "Sign language recognition by combining statistical DTW and independent classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 2040–2046, 2008.

[153] T. Arici, S. Celebi, A. S. Aydin, and T. T. Temiz, "Robust gesture recognition using feature pre-processing and weighted dynamic time warping," *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 3045–3062, 2014.

[154] A. Hernandez-Vela, M. Bautista, X. Perez-Sala, V. Ponce, X. Baro, O. Pujol, C. Angulo, and S. Escalera, "BoVDW: Bag-of-visual-and-depth-words for gesture recognition," in *ICPR*, 2012.

[155] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[156] M. Bhuyan, D. Ajay Kumar, K. MacDorman, and Y. Iwahori, "A novel set of features for continuous hand gesture recognition," *Journal on Multimodal User Interfaces*, vol. 8, no. 4, pp. 333–343, 2014.

[157] Y. Sun, N. Kuwahara, and K. Morimoto, "Development of recognition system of japanese sign language using 3D image sensor," in *HCI*, 2013.

[158] X. Suau, M. Alcoverro, A. Lopez-Mendez, J. Ruiz-Hidalgo, and J. Casas, "INTAIRACT: Joint hand gesture and fingertip classification for touchless interaction," in *ECCV*, 2012.

[159] C.-S. Lee, S. Chun, and S. W. Park, "Tracking hand rotation and various grasping gestures from an IR camera using extended cylindrical manifold embedding," *Computer Vision and Image Understanding*, vol. 117, no. 12, pp. 1711–1723, 2013.

[160] V. Hoang, A. Nguyen Hoang, and D. Kim, "Real-time stereo rendering technique for virtual reality system based on the interactions with human view and hand gestures," in *Virtual Augmented and Mixed Reality*, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469551, IEEE Transactions on Circuits and Systems for Video Technology

[161] K. O'Hara, G. Gonzalez, A. Sellen, G. Penney, A. Varnavas, H. Mentis, A. Criminisi, R. Corish, M. Rouncefield, N. Dastur, and T. Carrell, "Touchless interaction in surgery," *Commun. ACM*, vol. 57, no. 1, pp. 70–77, 2014.

[162] M. Strickland, J. Tremaine, G. Brigley, and C. Law, "Using a depth-sensing infrared camera system to access and manipulate medical imaging from within the sterile operating field," *Canadian Journal of Surgery*, vol. 56, no. 3, pp. E1–E6, 2013.

[163] J. Hoey, P. Poupart, A. von Bertoldi, T. Craig, C. Boutilier, and A. Mihailidis, "Automated handwashing assistance for persons with dementia using video and a partially observable markov decision process," *Computer Vision and Image Understanding*, vol. 114, no. 5, pp. 503–519, 2010.

[164] H. Cheng, Z. Liu, Y. Zhao, G. Ye, and X. Sun, "Real world activity summary for senior home monitoring," *Multimedia Tools and Applications*, vol. 70, no. 1, pp. 177–197, 2014.

[165] P. Yanik, J. Manganelli, J. Merino, A. Threatt, J. Brooks, K. Green, and I. Walker, "Use of Kinect depth data and Growing Neural Gas for gesture based robot control," in *International Conference on Pervasive Computing Technologies for Healthcare*, 2012.

[166] A. Dragan and S. Srinivasa, "Formalizing assistive teleoperation," in *Robotics: Science and Systems*, 2012.

[167] D. Fischinger, P. Einramhof, K. Papoutsakis, W. Wohlkinger, P. Mayer, P. Panek, S. Hofmann, T. Koertner, A. Weiss, A. Argyros, and M. Vincze, "Hobbit, a care robot supporting independent living at home: First prototype and lessons learned," *Robotics and Autonomous Systems*, In Press, 2014.

[168] G. Amayeh, G. Bebis, A. Erol, and M. Nicolescu, "Hand-based verification and identification using palm-finger segmentation and fusion," *Computer Vision and Image Understanding*, vol. 113, no. 4, pp. 477–501, 2009.

[169] V. Kanhangad, A. Kumar, and D. Zhang, "Contactless and pose invariant biometric identification using hand surface," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1415–1424, 2011.

[170] M. Roccetti, G. Marfia, and A. Semeraro, "Playing into the wild: A gesture-based interface for gaming in public spaces," *Journal of Visual Communication and Image Representation*, vol. 23, no. 3, pp. 426–440, 2012.

[171] S. Waldherr, R. Romero, and S. Thrun, "A gesture based interface for human-robot interaction," *Autonomous Robots*, vol. 9, pp. 151–173, 2000.

[172] X. Yin and X. Zhu, "Hand posture recognition in gesture-based human-robot interaction," in *IEEE Conference on Industrial Electronics and Applications*, 2006.

[173] X. Yin and M. Xie, "Finger identification and hand posture recognition for human-robot interaction," *Image and Vision Computing*, vol. 25, no. 8, pp. 1291–1300, 2007.

[174] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human-robot interaction," *Image and Vision Computing*, vol. 25, no. 12, pp. 1875–1884, 2007.

[175] D. Michel, K. Papoutsakis, and A. Argyros, "Gesture recognition for the perceptual support of assistive robots," in *International Symposium on Visual Computing*, 2014.

[176] F. Pedersoli, S. Benini, N. Adami, and R. Leonardi, "XKin: an open source framework for hand pose and gesture recognition using kinect," *The Visual Computer*, vol. 30, no. 10, pp. 1107–1122, 2014.

[177] P. Morerio, L. Marcenaro, and C. Regazzoni, "Hand detection in first person vision," in *International Conference on Information Fusion*, 2013.

[178] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Cucchiara, "Hand segmentation for gesture recognition in EGO-vision," in *ACM IMMPD*, 2013.

[179] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *IEEE CVPR*, 2013.

[180] ——, "Model recommendation with virtual probes for egocentric hand detection," in *IEEE ICCV*, 2013.

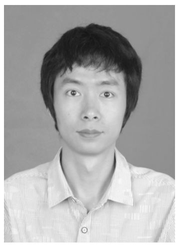
[181] A. Betancourt, M. Lopez, C. Regazzoni, and M. Rauterberg, "A sequential classifier for hand detection in the framework of egocentric vision," in *IEEE CVPR Workshops*, 2014.

[182] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara, "Gesture recognition in ego-centric videos using dense trajectories and hand segmentation," in *IEEE CVPR Workshops*, 2014.

[183] F. Coleca, T. Martinetz, and E. Barth, "Gesture interfaces with depth sensors," in *Time-of-Flight and Depth Imaging*, 2013.

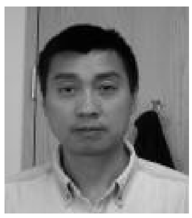
Hong Cheng (M'06-SM'14) received the Ph.D. degree in Pattern Recognition and Intelligent Systems from Xi'an Jiaotong University, Xi'an, China, in 2003. He is currently a full Professor in School of Automation Engineering, University of Electronic Science and Technology of China (UESTC). He is the executive director of Center for Robotics, UESTC. Before this, He was a postdoctoral researcher at School of Computer Science, Carnegie Mellon University, USA from 2006 to 2009 and an associate Professor of Xi'an Jiaotong University since 2005.

Dr. Cheng is a senior member of IEEE, ACM, and Associate Editor of IEEE Computational Intelligence Magazine. He is a reviewer of many important journals and conferences (IEEE TITS, MAV, CVPR, ICCV, ITSC, IVS, ACCV, etc.). Dr. Cheng serves as the General Chair of VALSE 2015, Finance Chair of ICME 2014, Local Arrangement Chair of VLPR 2012, Registration Chair of the ICVES 2005. He has over 80 academic publications including two books. His current research interests include computer vision, machine learning and robotics.



Lu Yang (M'11) received the Ph.D. degree in Electrical Engineering and Computer Science from Nagoya University, Nagoya, Japan, in 2011. From 2008 to 2011, he was sponsored by China Scholarship Council (CSC) to study in Nagoya University. From 2005 to 2008, he was in the joint doctor-master program in the School of Automation Engineering, University of Electronic Science and Technology of China (UESTC), where he also received the Bachelor degree in 2005. He came back UESTC in 2011 and joined the School of Automation Engineering as a Lecturer. Since 2013, he has been an associate Professor in the Center for Robotics of UESTC.

Dr. Yang is the member of IEEE and committee member of ICME 2014, VALSE 2015, ChinaSIP 2015 and PCM 2015. His current research interests include computer vision, pattern recognition, machine learning and 3D multimedia applications.



Zicheng Liu (SM'05-F'15) received the BS degree in mathematics from HuaZhong Normal University, Wuhan, China, in 1984, the MS degree in operation research from the Institute of Applied Mathematics, Chinese Academy of Sciences in 1989, and the PhD degree in computer science from Princeton University in 1996. He is a principle researcher at Microsoft Research Redmond. Before joining Microsoft Research, he was at Silicon Graphics Inc.

Dr. Liu is a fellow of IEEE. His current research interests include human activity recognition, 3D face modeling and animation, and multimedia signal processing.