A Review of Constraints on Vision-based Gesture Recognition for Human-Computer Interaction

Biplab Ketan Chakraborty[1], *Debajit Sarma*[1], *M.K. Bhuyan*[1*], *Karl F. MacDorman*[2]

[1] *Department of Electronics and Electrical Engineering, IIT Guwahati, India 781039.*
[2] *Indiana University School of Informatics and Computing, 535 West Michigan St., Indianapolis, IN 46202 USA.*
** E-mail: mkb@iitg.ernet.in*

Abstract: The ability of computers to recognize hand gestures visually is essential for progress in human–computer interaction. Gesture recognition has applications ranging from sign language to medical assistance to virtual reality. However, gesture recognition is extremely challenging not only because of its diverse contexts, multiple interpretations, and spatio-temporal variations but also because of the complex non-rigid properties of the hand. This paper surveys major constraints on vision-based gesture recognition occurring in detection and pre-processing, representation and feature extraction, and recognition. Current challenges are explored in detail.

1 Introduction

The use of computers and related devices has become ubiquitous. Hence, the need has increased for interfaces that support effective human–computer interaction (HCI). *HCI* concerns "the design, evaluation, and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them" [1, 2], especially "in the context of the user's task and work" [3]. In striving toward these ends, HCI research builds on progress in several related fields, as shown in Fig. 1. It is focused not only on enhancing the usability, reliability, and functionality of present-day interfaces but also on the development of novel, innovative interfaces that can be used in natural, lifelike ways. Such interfaces are in demand for interacting with virtual environments in computer games and virtual reality, for teleoperation in robotic surgery, and so on. Active interfaces, intelligent adaptive interfaces, and multimodal interfaces are all gaining prominence.

Gesture recognition is an important area for development in HCI systems. These systems can be broadly classified based on their number of channels as *unimodal* or *multimodal* [4]. Unimodal systems can be a) *vision-based* (*e.g.*, body movement tracking [5], facial expression recognition [6, 7], gaze detection [8], and gesture recognition [9]), b) *audio-based* (*e.g.*, auditory emotion recognition [10], speaker recognition [11], and speech recognition [12]), or c) based on other types of sensors [13]. The most researched unimodal HCI systems are vision based. People typically use multiple modalities during human–human communication. Therefore, to assess a user's intention or behaviour comprehensively, HCI systems should integrate information from multiple modalities as well [14]. Multimodal interfaces can be setup using combinations of inputs, such as gesture and speech [15] or facial pose and speech [16]. Most of the recent and important applications of VGR include Sign language recognition [17], Virtual reality [18], Virtual game [19], Augmented reality [20], Smart video conferencing [21], Smart home and office [22], Healthcare and Medical assistance (MRI navigation) [23], Robotic surgery [24], Wheelchair control [25], Driver monitoring [26], Vehicle control [27], Interactive presentation module [28], and Virtual classroom [29], e-commerce [30], and so on.

In recent years, vision-based gesture recognition (VGR) has become a key research topic in HCI and there are many real-life applications of VGR. One of the breakthroughs in VGR is the introduction of Microsoft Kinect® as a contact-less interface [31]. The Kinect has significant potential in various applications, such as healthcare [23], education [32], etc. However, its poor outdoor performance and depth resolution limit its usability. Recently, SoftKinetic's Gesture Control Technology is incorporated in BMW cars to allow drivers to navigate the in-vehicle infotainment system with ease [33]. However, the VGR module is still in its primitive stage due to its limited functionality and gesture vocabulary. Although VGR has been surveyed in several review papers (Table 1), none specifically addresses problems arising from constraints on existing VGR methods. The nature and extent of these constraints depends on the application. Major difficulties associated with image processing techniques during gesture recognition for the above mentioned VGR applications are as follows:

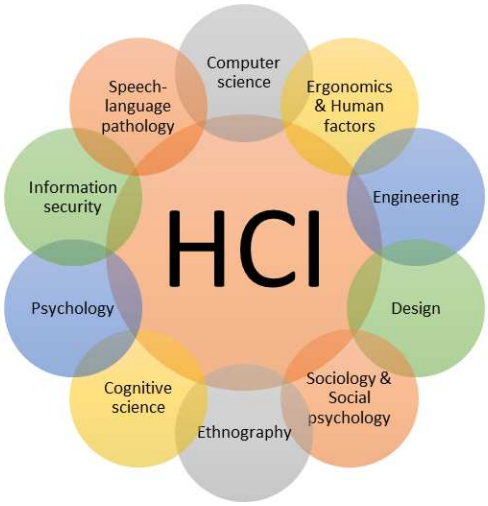

Fig. 1: Human-computer interaction and related research fields

- A gesture recognition system generally uses a single colour video camera to minimize the necessary hardware components and calibrations. Using a single colour video camera and consequently the projection of the 3D scene on a 2D plane results in the loss of depth information. Therefore, reconstruction of the 3D trajectory of the hand in space is not possible.
- Illumination change in the testing environment seriously hampers hand image segmentation.
- For reliable recognition, hand image segmentation via background motion compensation and accurate motion estimation of the hand is

This is the author's manuscript of the article published in final edited form as:
Chakraborty, B. K., Sarma, D., Bhuyan, M. K., & MacDorman, K. F. (2018). Review of constraints on vision-based gesture recognition for human–computer interaction. IET Computer Vision, 12(1), 3–15. https://doi.org/10.1049/iet-cvi.2017.0052

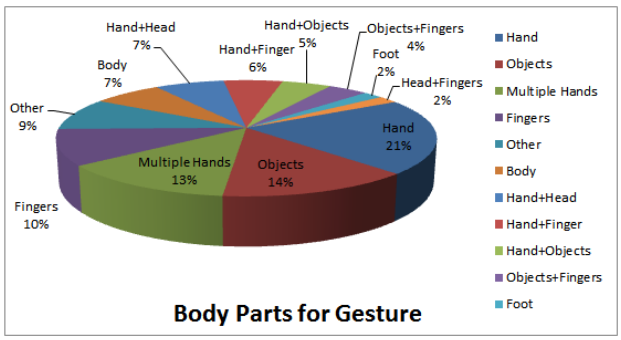

Fig. 2: Use of the different body parts or objects for gesturing [45]

an essential requirement. Unfortunately, like other image processing problems, hand tracking and segmentation in a cluttered background is a critical problem in hand gesture recognition. In most occasions, the background is not simple. Also, the background may contain other body parts captured by the camera. Moreover, the occlusion due to other body parts or self occlusion may cause problems during segmentation.

- The processing of a large amount of image data is time consuming and thus, real- time recognition may be difficult.

However, some of the applications mentioned above require more attention towards some constraints than others, as highlighted below:

- VGR systems in 3D modelling-based applications such as virtual and augmented reality require accurate depth estimation with high depth resolution.
- VGR systems for indoor applications such as sign-language recognition and virtual reality require uniform ambient illumination. However, VGR systems for outdoor applications such as driver monitoring, vehicular control, etc. should work in a environment with a broad range of dynamically changing illuminations and cluttered backgrounds.
- Applications requiring real-time multiple user interaction such as smart video conferencing, smart home and office, virtual classrooms, e-commerce, etc. require a properly synchronized hardware setup for real-time applications.
- Higher recognition accuracy is a primary criterion for VGR applications like healthcare and medicine. In addition, portability is important for some VGR applications, such as vehicles and other mobile platforms.

Section 2 presents an overview of a basic VGR system. Section 3 examines main constraints on gesture recognition systems for vision-based HCI modules.

2 Overview of vision-based gesture recognition

Gesture recognition is an important area of research in computer vision. In the framework of human–computer non-verbal communication, the visual interface establishes communication channels for inferring intentions from human behaviour, including facial expressions and hand gestures. Hand gesture recognition from visual images forms a key part of this interface. It has potential applications in machine vision, virtual reality, robotic control, and so on.

2.1 Gesture recognition system architecture

A gesture is a pose or physical movement of the hands, arms, face, or body that is meaningful and informative. According to Karam [45], the hand is the most widely used body part (Fig. 2). The primary task of vision-based interfaces (VBI) is to detect and recognize visual information for communication. A vision-based approach is more natural and convenient than, for example, a glove-based approach. It is easy to deploy and can be used anywhere within a camera's field of view. However, it is more difficult because of current limitations in computer vision in processing human hands, which are non-convex and flexible. The straightforward approach to vision-based gesture recognition is to acquire visual information about a person in a certain environment and try to extract the necessary gestures. This approach must be performed in a sequence, namely, detection and pre-processing, gesture representation and feature extraction, and recognition.

1. *Detection and pre-processing:* The detection of gesturing body parts is crucial because the accuracy of the VGR system depends on it. Detection includes capturing the gestures using imaging devices. Pre-processing segments the gesturing body parts from images or videos as accurately as possible given such constraints as illumination variation, background complexity, and occlusion.
2. *Gesture representation and feature extraction:* The next stage in a hand gesture recognition task is to choose a mathematical description or model of the gesture. The scope of a gestural interface is directly related to the proper modelling of hand gestures. Modelling depends on the intended application. For some applications, a coarse, simple model may be sufficient. However, if the purpose is natural interaction, a model should allow many, if not all, natural gestures to be interpreted by the computer. Many authors have identified different features for representing particular kinds of gestures [46]. Most of the features can be broadly classified as follows: a) shape, (*e.g.*, geometric features or nongeometric features [34]), b) texture or pixel value, c) 3D model-based features [44], and d) spatial features (*e.g.*, position and motion [47–49]).
3. *Recognition:* The final stage of a gesture recognition system is classification. A suitable classifier recognizes the incoming gesture parameters or features and groups them into either predefined classes (supervised) or by their similarity (unsupervised) [9]. There are many classifiers used for both static and dynamic gesture recognition, each with its own limitations.

3 VGR systems and their associated constraints

The ability of computers to recognize hand gestures visually is essential for the future development of HCI. However, vision-based recognition of hand gestures, especially dynamic hand gestures, poses a onerous difficult interdisciplinary challenge for three reasons:

- Hand gestures are diverse, have multiple meanings, and vary spatio-temporally;
- The human hand is a complex non-rigid object making it difficult to recognize; and
- Computer vision itself is an ill-posed problem.

A gesture recognition system relies on a series of subsystems, as explained in the last section. Because the subsystems are connected in series, the overall accuracy of the system is the product of the accuracy of each subsystem. Thus, overall performance cannot exceed that of the subsystem that is the "weakest link". This section presents an analysis of the difficulties associated with hand gesture recognition at each stage (Fig. 3).

3.1 Detection and pre-processing

Gesture detection involves capturing images or videos using imaging devices and then, in a pre-processing stage, localizing in them the gesturing body parts.

3.1.1 Gesture acquisition: The acquisition subsystem contains either a 2D camera (b/w or colour) or a 3D imaging system (*e.g.*, a stereo camera pair or a camera with depth sensors). A camera consists of several subsystems. Berman and Stern [50] reviewed the effects of camera parameters on VGR system characteristics. VGR accuracy depends on the following camera specifications:

Table 1 Some comprehensive surveys on gesture recognition and gesture-based interfaces

Ref.	Scope of analysis	Contributions
Pavlovic *et al.* [34] (1997)	A survey of hand gesture modelling, analysis, and recognition.	3D hand models better describe and discriminate gestures but with higher computational complexity. Appearance-based models are limited in gesture description despite being simpler to implement.
Ying *et al.* [35] (1999)	A survey of vision-based static hand posture and temporal gesture recognition approaches and their applications.	An investigation of feature selection, data collection for visual gesture learning, various algorithms for static and dynamic gesture recognition, use of HMMs, and their variants in sign language recognition.
Moeslund *et al.* [36] (2001)	A survey of vision-based human motion capturing methods focusing on initialization, tracking, pose estimation, and recognition.	Investigation of different problems such as the lack of an alphabet vocabulary of motion data for training, the large amount of time required to capture and label human motion data, use of 2D silhouettes to estimate a 3D pose, problem of incremental updates, motion ambiguity due to movement projection on planes.
Derpanis *et al.* [37] (2004)	A review of vision-based hand gestures on different aspects like the feature set, classification methods and gesture representation.	Analysis of gesture-based HCI under ambient lighting conditions and high contrast stationary backgrounds.
Moeslund *et al.* [5] (2006)	A review of advances in video-based human capture and analysis and open problems in the automatic visual analysis of human movement.	Investigated automatic initialization of model shape, appearance, and pose; reliable detection and tracking of multiple people in unstructured outdoor scenes with partial occlusion; human motion reconstruction from multiple views and from monocular videos.
Erol *et al.* [38] (2007)	A review of hand pose estimation methods and capturing of the real 3D hand motion in HCI.	Comparison of methods on the effective number of DoF, number and type of cameras, system's operation in a cluttered background, use of hand model constraints, algorithmic details, execution speed, and observable pose restrictions. Discussion of problems due to lack of ground-truth data, high-dimensionality, self-occlusions, uncontrolled environments, and rapid hand motion.
Mitra *et al.* [9] (2007)	A survey of hand gesture and facial expression recognition methods with discussion on different tools for gesture recognition.	Discussion on HMMs, particle filtering, and condensation algorithm, FSMs, ANNs, Gabor filtering and optical flow; hybridization of HMMs and FSMs; different facial expression modelling approaches; use of soft computing tools.
Moni *et al.* [39] (2009)	A literature review of different approaches to vision-based gesture recognition using HMMs.	Difficulty in achieving a continuous online recognition with minimal time delay because the starting and ending point of a gesture are not known. Problems due to assumptions about lighting, background, and signer's location are discussed.
Wachs *et al.* [40] (2011)	A description of the requirements of hand-gesture based interfaces for various application types and associated challenges.	Discussions on gesture recognition methods that depend on the application, domain, environment, and the user's cultural background.
Suarez *et al.* [41] (2012)	A review of depth image-based hand tracking and gesture recognition systems.	Investigation on the use of Kinect for hand segmentation using depth thresholding; hand tracking using Kalman filters and mean shift clustering or using the NITE body-tracking and hand-tracking module from OpenNI.
Rautaray *et al.* [42] (2012)	A comparative survey of hand gesture interfaces, focusing on gesture recognition systems based on detection, tracking, and recognition, and their applications.	Comparative analysis on gesture taxonomies, gesture representation and recognition techniques, different software platforms. Also, a survey on application domain of gesture recognition is provided.
Hasan *et al.* [43] (2013)	A survey of gesture-based HCI focusing on different application domains for efficient interaction.	Analysis focused on different applications of gesture recognition techniques. Investigation also found appearance-based gesture representations are preferable than 3D-based gesture representations due the complexity of implementation of the latter.
Cheng *et al.* [44] (2016)	A survey of various 3D hand gesture modelling and recognition techniques.	Investigation on different 3D hand gesture modelling and recognition techniques, *e.g.*, static hand gesture, trajectory-based hand gesture and continuous hand gesture. Also, discuss different 3D depth sensors along with some popular dataset.

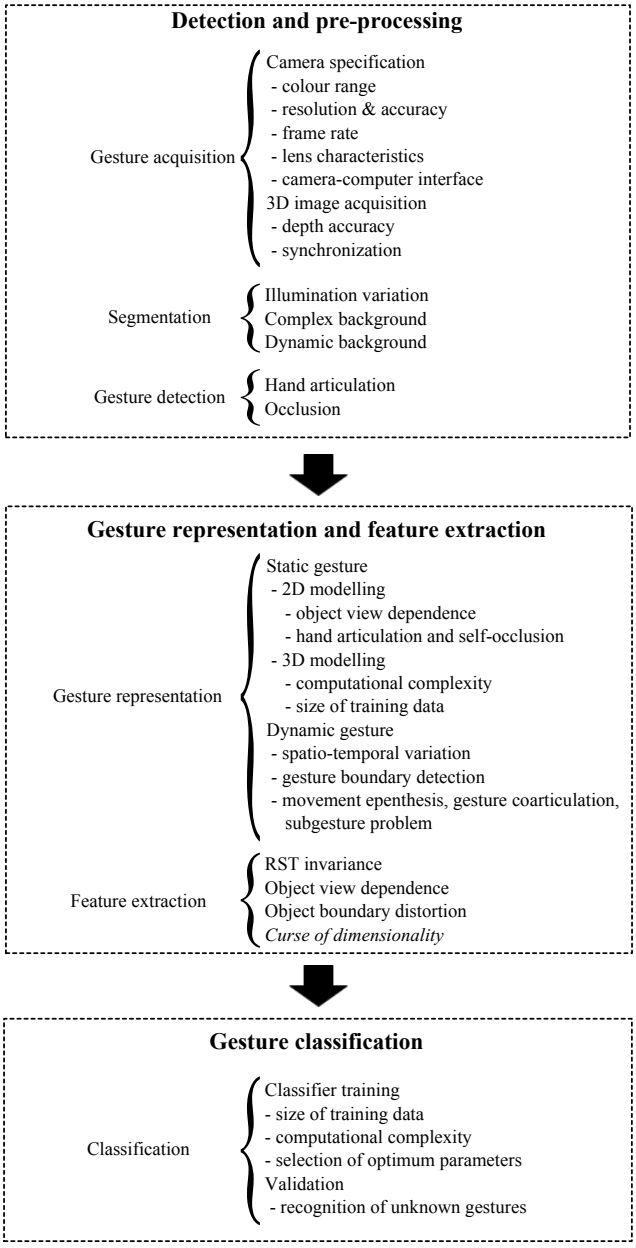

Fig. 3: Different stages of a VGR system and their associated constraints.

1. *Colour range:* For image acquisition, sensors for the visible light range (390–750 nm wavelength) are typically used. The main problem with visible light sensors is that their response depends on the amount and chromatic nature of an illuminant. Near-infrared (IR) radiation offers an alternative to visible light [51] that is less sensitive to illumination changes. Human gestures can be captured using near-IR cameras with subjects wearing near-IR reflectors. However, the reflectors may impair the user's natural feeling of gesturing. By contrast, thermal or infrared (IR) cameras depend solely on the heat signature generated by the human body. Thus, IR cameras are superior to visible light cameras under poor or varying illumination conditions. However, visible cameras are usually preferred because of their low cost.
2. *Resolution and accuracy:* The picture quality of a camera depends on two factors: its resolution and its colour reproduction accuracy. The resolution of a camera is represented by the number pixels present in a column (*e.g.*, 720 pixels, 1080 pixels), by both the column and row (*e.g.*, 640×480 pixels), or by the total number pixels (*e.g.*, 0.3 megapixels). Accuracy describes how accurate the colour reproduction is. So, each pixel value is represented by several bits (*e.g.*, 8 bits, 24 bits). Standard RGB cameras have 24-bit accuracy. According to Zhenyao and Neumann [52], use of low resolution cameras may affect the performance of a VGR system in the following ways:
- The segmentation of the hand may be noisy because of a complex background or varying illumination conditions.
- For applications like sign language recognition, a full upper-body image is required, and so hand regions occupy a tiny area in the image. Hence, it is difficult to extract texture and other geometrical features from the hand.
- A VGR system may fail to initialize or track gestures at a low resolution.

Use of high resolution cameras (HD cameras) can resolve the aforementioned constraints at the expense of processing time.
3. *Frame rate:* The number of images or *frames per second* (fps) a camera can capture. In a VGR system, the frame rate should be enough to capture body part movements accurately. The most appropriate frame rate depends on the hardware configuration of the image acquisition system and on the speed of gesturing hands or other body parts. For slowly moving hands, a low frame rate such as 15 fps may be satisfactory. However, applications such as sign language recognition or virtual reality involve fast moving hands or body parts. In these cases, a low frame rate camera may not able to capture enough instances of moving hands for continuous motion. Some of the existing frame rate standards are 24 fps (standard film), 25 fps (PAL), 29.97 fps (NTSC), and 60 fps (ATSC video).
4. *Lens characteristics:* The quality and amount of distortion present in an image are influenced by camera lens characteristics. A lens is characterized by its focal length (in mm) and aperture (F-number). The focal length describes the field of view (FoV) of a camera. The aperture is the amount of opening, which determines the depth of field, that is, the range of distances at which objects appear in focus. The depth of field (DoF) reduces with an increasing aperture for a lens with a fixed focal length. It results in the blurring of distant objects with an increasing F-number. Most of the image acquisition systems use cameras having fixed lenses and aperture. Ideally, the performance of a VGR system should not be affected by illumination variations and it should have a large DoF. A camera having a narrow aperture gives a large DoF, but less light enters the camera. This results in noisy and dark images under low light conditions. Therefore, hand detection and segmentation could be difficult. By contrast, cameras with a larger aperture provide a better quality image under low light conditions but with fewer DoF. Therefore, a trade-off exists between image brightness and DoF.
5. *Camera-computer interface:* The data transfer speed of the interface between a camera and a computer significantly affects the overall speed of a VGR system. For a video of resolution $W \times H$ with frame rate f_r, the required bitrate for transmission in uncompressed form is given by,

$$bitrate = 24 \times f_r \times W \times H$$

So, a 640×480 (30 fps) video requires ~ 211 Mb/s data transfer rate, whereas a high definition video (720p, 30 fps) requires a data transfer rate ~ 633 Mb/s. Now, most of the low end cameras use a USB 2.0 interface with a maximum speed limit of 480 Mb/s. However, the maximum achievable data transfer rate depends on the total number of USB peripherals connected to the system. This limits the data transfer rate to a lower value which restricts the resolution and frame rate of the video to be transmitted. Also, the computer's run time load and capabilities could degrade the resolution and frame rate supported by the camera. In recent years, some high-end cameras use a USB 3.0 interface (4.8 Gb/s) equipped with built in codecs. This reduces the bitrate burden to the interface by compressing the data before transmission.

Although most VGR systems use 2D cameras, the focus of recent research is shifting towards incorporating depth information. The use of 3D sensors in a VGR system improves its overall robustness and flexibility while increasing its cost and computational complexity. The 3D information can be used either to model a 3D gesture by

directly using the 3D position vectors or as a frontal plane position (colour information) plus range (depth information). For the latter, range information is used mostly to segment the foreground (gesturing body parts) from the background. The colour information is used for 2D gesture recognition. The different methods of data acquisition in 3D VGR systems are as follows:

1. *Standard stereo cameras:* A pair of standard colour video or still cameras capture two simultaneous images. Disparity is calculated as a depth measure from the two simultaneous images. Some of the gesture recognition methods using stereoscopic systems are given in [53, 54]. A major challenge in stereoscopic vision is the finding correspondence between a stereo image pair. Difficulties arise for many reasons:
(a) *Camera inconsistency:* The geometrical configuration (intrinsic parameters) of two cameras are not the same, and even the parameters of two cameras of same brand are not identical. So, there is inconsistency between the two captured images. Presence of noise in any one or both of the images changes the pixel intensity values, which leads to improper matching.
(b) *Unique correspondence:* It is very difficult to find correspondence in the textureless image regions. Ambiguity in matching also occurs when the image regions have a similar or repetitive pattern or structure in the horizontal direction.
(c) *Occlusion:* In sign language recognition or two handed gesture recognition, one hand can occlude the other hand or face. So, finding proper correspondence between two images is difficult in the presence of occlusion.
Apart from this, accuracy of a disparity map may depend on other factors:
- The accuracy of disparity-based depth estimation is poor if the foreground and background are very close to each other. For example, a user could be found standing or sitting very near to a wall. In this case, the difference in disparity values of background and foreground regions could be small. This will result in a foreground-background ambiguity in the depth map.
- The disparity calculation is computationally expensive.
- The disparity calculation relies on adequate scene illumination and surface texturing. Also, the disparity map obtained from the stereo image pairs may not give information about the specular surface.

2. *Multiple camera-based system:* In contrast to the stereo camera-based acquisition system that gives only a relative depth measure between different objects and the camera, a multiple camera-based acquisition system can better capture the 3D structure of an object. In the last decade, several researchers [55, 56] used multiple camera-based systems for gesture recognition. Although most of the multiple camera-based VGR systems used visible light cameras, some of the researchers used IR cameras. For example, Ogawara *et al.* fitted a 26-degree of freedom (DoF) kinematic model to a volumetric model of the hand constructed from images obtained using multiple infrared cameras arranged orthogonally [57]. However, multiple camera-based gesture acquisition systems have the following limitations:
(a) *Synchronization:* The cameras should be synchronized to capture multi-view images of an object, which implies simultaneous capturing all multi-view images. However, all imaging devices have varying delays. Thus, a dedicated hardware setup is needed to synchronize all the cameras.
(b) *Ruggedness:* The system becomes complex and rigid. It is hard to mount a multiple camera-based system on a portable platform such as a notebook computer or smart phone. System performance depends so heavily on the device's structural configuration that, for even a small change in configuration, it requires recalibration.

3. *Time-of-flight (ToF) system:* This method obtains depth information by measuring the time delay between the transmissions and reception of a light pulse. Examples include
- PMD Technologies' CamCube®: depth range up to 60 m.
- Mesa Imaging's SwissRanger SR4000®: depth range of 5–10 m with 176×144 resolution.

Some recent gesture recognition methods using ToF are given in [58–60]. Though the inclusion of depth information can significantly improve the accuracy of gesture recognition, ToF systems face the following limitations:
- Time must be measured with a high accuracy. For example, an accuracy of ~ 1 cm in depth measurement requires an accuracy of ~ 66.7 picoseconds in time measurement. However, measurement of returned light pulse may be inexact due to light scattering near a VGR system. Thus, accurate depth information cannot be determined.
- It is difficult to generate short light pulses with fast rise and fall times.
- For a VGR system, the depth resolution should be high. The depth resolution of a ToF camera is proportional to its signal-to-noise ratio (SNR). The SNR is inversely proportional to the number of electrons collected, which depends on the amount of reflected light [61]. A light source having a low pulse repetition rate (*e.g.,* lasers) produces low SNR, thus yielding low depth resolution.
- Background light and multiple reflections create interference and cause ambiguity in time measurement, which results in inaccurate depth measurement.

According to Basler's ToF sensor manual [62], the following constraints should be imposed on hardware installation to avoid background light/scatterings:
- the camera should not be exposed to bright sunlight (ambient illumination < 15 klx)
- the presence of mirrors or any other reflective surfaces/objects should be avoided in the vicinity of the sensor to reduce scattering.

4. *Light coding technology:* This method uses an infrared projector to project a known pattern on the object and a CMOS sensor to capture deformations in the reflected pattern. Depth information is then calculated from the amount of deformation. Examples include
- PrimeSense®: single-chip based, and depth image resolutions up to 640×480 pixels (minimum distance ~ 0.8 m).
- Microsoft Kinect®: depth range up to 5 m (minimum distance ~ 1.8 m).

Some recent gesture recognition methods using Microsoft Kinect are given in [63, 64]. The major limitations of light coding systems are
- Poor performance in outdoor environments.
- The error in depth measurements increases quadratically with distance from the sensor (about 4 cm at the distance of 5 m).
- Depth resolution decreases quadratically with the distance from the sensor.

5. *Leap motion sensors:* The objective is to determine 3D fingertip positions instead of whole body depth information as with the Kinect sensor. The sensor can obtain these positions with high accuracy (~ 0.01 mm) [44]. Leap motion sensors can be used in various applications, *e.g.*, virtual environments [65] and gesture [66] and sign language recognition [67]. The major limitations of leap motion sensors are
- The sensors use specialized hardware and software and, thus, may be less compatible with other systems.
- The sensor detects only fingertips lying parallel to the sensor plane. So, it may fail to detect multiple fingertips when fingers overlap (*e.g.*, sign language). This restricts its applicability in the recognition of complex gestures.
- The system may lack accuracy in normal lighting conditions or fail to track hands or fingers during an activity.

3.1.2 Segmentation: Accurate segmentation of hand or body parts from the captured images remains a challenge in computer vision for many reasons. Some of the limitations in segmentation are as follows:

1. *Illumination variation:* Illumination variation affects the accuracy of skin colour segmentation methods. Poor illumination may change the chrominance properties of the skin colours, and the skin colour will be different from the image colour as shown in Fig. 4. Several researchers [68–70] converted from RGB to another colour space, dropped the luminance component, and used only the chrominance components to compensate for brightness variations in the image. However, because the skin reflectance locus and the illuminant locus are directly related [71], the perceived colour depends on scene illumination. Biplab *et al.* [72, 73] used a fusion-based image

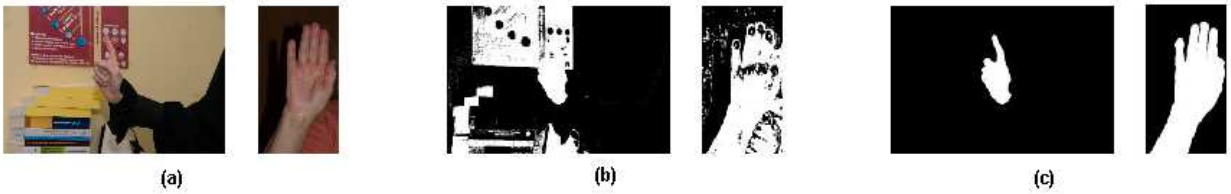

Fig. 5: Effect of complex background on skin colour segmentation: (a) original images, (b) segmentation results, and (c) ground truth

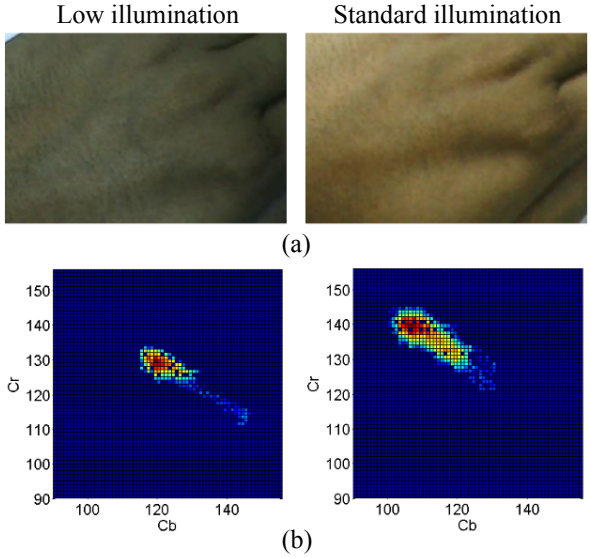

Fig. 4: Effect of illumination variations on perceived skin colour: (a) Skin colour in low and standard illumination conditions, (b) 2D colour histogram in YCbCr space

specific model for skin segmentation under varying illumination conditions.

2. *Complex background:* A major challenge in gesture recognition is the proper segmentation of skin coloured objects (*e.g.,* hands, face) against a complex static background. The accuracy of skin segmentation algorithms is limited because of objects in the background that are similar in colour to human skin (Fig. 5). Skin colours in the background increase false positives. This problem can be addressed by different approaches such as using additional features, *e.g.,*, texture features [74, 75]. This assumes that a sharp textural discrimination exists between skin and non-skin regions. That assumption, however, may be violated due to the low resolution of images and/or a smooth background.

3.1.3 Hand articulation and occlusion: The hand is an articulated object with more than 20 DoF [76]. Many hand parameters need to be estimated for different hand poses, locations, and orientations. Their, estimation is complicated because of the hand's high DoF. For monocular vision, the occlusion of joints and finger segments can prevent the detection of a hand configuration. Applications in vision-based interfacing (VBI) need to work around these limitations by using gestures that do not require full hand pose information. So far, articulated hand pose estimation in unconstrained settings is a largely unsolved problem.

Mitigating the effects of occlusion is a major challenge for gesture recognition. Not only may the hand occlude itself, but one hand may occlude the other during two-handed gestures. Both kinds of occlusion affect the appearance of the hand, thus hindering gesture recognition. In monocular vision-based gesture recognition, the appearance of gesturing hands is viewpoint dependent. As shown in Fig. 6, different hand poses appear similar from a particular viewpoint because of self-occlusions.

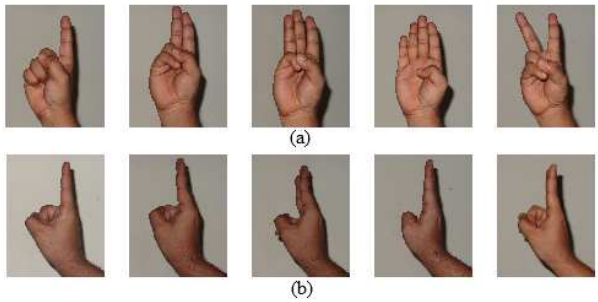

Fig. 6: Different hand poses and their side views

3.2 Gesture representation and feature extraction

To recognize a gesture, it must be represented using a suitable model. According to its spatio-temporal properties, gestures are broadly classified as *static* or *dynamic*. *Static* gestures are defined by the pose or orientation of a body part in the space (*e.g.*, hand pose), whereas *dynamic* gestures are defined by the temporal deformation of body parts (*e.g.*, shape, position, motion). After gesture modelling, an important step in gesture recognition is the extraction and selection of suitable features to represent associated gestures. A discussion follows on different gesture representation techniques and feature descriptors, and their constraints:

3.2.1 Gesture representation: Static gestures can be represented using a 2D model or a 3D model as shown in Fig. 7. The simplest way to represent a static gesture or pose is to use an appearance-based model, which tries to identify gestures directly from visual images. The model parameters are not derived from a 3D spatial description but instead by relating the appearance of the gesture to that of a set of predefined template gestures. Parameters of such models may be either the images themselves or some features derived from the images. Appearance-based models directly link the appearance of the hand and arm movements in visual images to specific gestures. The application of 2D modelling of static hand gestures is limited because it is object view dependent. 2D models are not sufficient to represent complex static hand gestures (*e.g.*, sign language) because of the hand's articulation. Recent work focuses on the 3D representation of articulated hand gestures and their recognition. 3D static hand gesture recognition methods can be broadly classified into two model-based approaches: discriminative and generative. Discriminative approaches do not explicitly generate a 3D model of hands or other body parts. Instead, the classifiers are trained with appearance-based hand features to map unknown hand shape parameters with appearance-based features. This requires a large set of training data for offline training. In recent years, several researchers used 3D discriminative models for static hand gesture recognition. Shotton *et al.* [77] used skeletal modelling of body parts using depth data derived from a Kinect sensor. Keskin *et al.* [78] used a 3D mesh model to generate synthetic hand pose images to train randomized decision trees (RDTs). However, depth features are sensitive to background changes. To mitigate this problem, Yao and Fu [79] fused different features (*e.g.*, shape, surface, and position features) to use in a random forest (RF) classifier. The main limitations of discriminative models are

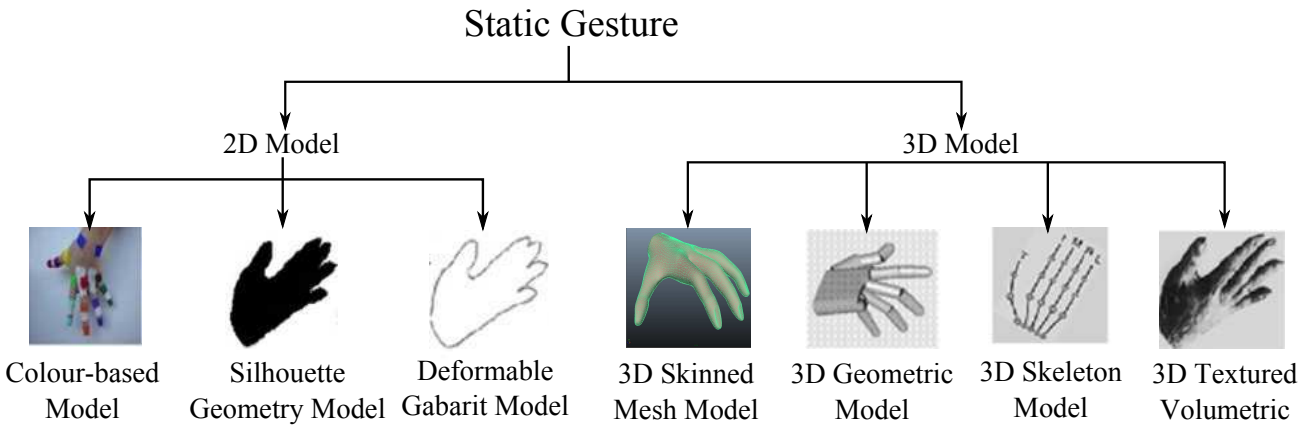

Fig. 7: Different hand models for static VGR system

- The need for a large amount of data to train the classifiers, and
- A lack of robustness to real-life problems like hand articulation and self-occlusions due to the absence of kinematic constraints on hand modelling.

In contrast to discriminative modelling, generative modelling tries to fit 3D models explicitly to the data obtained from 2D or 3D images. The 3D hand appearance can be approximated using geometric structures like cylinders, spheres, ellipsoids, and hyper-rectangles [34]. 3D hand modelling can be performed using a kinematic model with 28 DoF [80] derived from 2D silhouette images. Unfortunately, silhouette images are not sufficient to model real-life challenges like self-occlusion or varying illumination. De LaGorce and Paragios [81] showed that the use of texture mapping and a shading model can mitigate these problems. Researchers like Oikonomidis *et al.* used multiple camera-based images [82] or RGB-D images from a Kinect sensor [83] to match the 26-DoF hand model. In these methods, the 3D hand surface is represented with basic 3D shapes like a sphere and multiple truncated cylinders. The main limitations of generative models are their

- Computational intensiveness, and
- Lack of portability owing to the need for special hardware, like a multiple camera system.

As compared with static gesture recognition, which only uses features like colour, texture, and shape, dynamic gesture recognition uses motion and/or deformation information like position, skewness, and velocity. Dynamic gestures can be classified as isolated gestures or continuous gestures:

1. *Isolated gesture:* These gestures are temporally discrete. Their representation can be subdivided into two categories of feature descriptors: appearance-based and tracking-based. Gesture representation with appearance-based features mostly relies on the local feature descriptor. Dynamic features include the velocity and acceleration of a hand. The features can be obtained either from 2D or 3D image sequences (*e.g.*, RGB-D). Bhuyan *et al.* proposed both static and the dynamic features for trajectory guided recognition of hand gestures from the sequence of 2D frames [47–49]. The main problem with trajectories made of 2D data is their susceptibility to occlusions. Recent advancements in 3D imagery shifted the focus of researchers from 2D to 3D trajectories. Wang *et al.* [53] used relative depth information, hand displacement, and concatenated body centroids to model gesture appearance. In another work [84], they used random occupancy pattern features derived from depth information to handle the effects of occlusion. Other researchers [85, 86] used a concatenated histogram of oriented gradients (HOG) feature descriptor derived from 3D data to represent the hand gesture.

In contrast to appearance-based features, tracking-based features need explicit tracking of the skeleton of the gesturing body parts and corresponding centroids. According to the findings of Sohn *et al.* [87], frequently used parameters for hand tracking include hand centroid position, velocity, acceleration, and chain code. The hand centroid location can either be obtained by segmentation [88] or 3D body skeleton [89]. Features like hand orientation can also be used for 3D trajectories of hands [90].

All the above features are then fed into a tracking algorithm, such as a Kalman filter [91] or a particle filter [92]. Kalman filtering is an optimized tracking technique employed to estimate changes over time. Its limitations include imprecision of the system model and feature measurements and probability distributions limited mainly to the Gaussian. Another powerful algorithm, the condensation algorithm (conditional density propagation), also called particle filter or Monte-Carlo algorithm, allows general representations of probability, given that objective states are often non-Gaussian. This enables the use of nonlinear motion models, such as factored sampling, to approximate arbitrary probability distributions. Though the condensation algorithm is robust, particularly to background clutter, it has the disadvantage that the object model must be known in advance. Moreover, the iterative process is still prone to flounder at some step.

2. *Continuous gesture:* Applications using the human hand as a human-computer interface motivate researchers on continuous hand gesture recognition. For sign language, the recognition engine must support natural gesturing to enable the user's unrestricted interaction with the system. Because non-gestural movements often intersperse a gesture sequence, these movements should be removed from the video input before the gesture sequence is identified. Examples of non-gestural movements include movement epenthesis—the movement that occurs between gestures—and gesture coarticulation—the effect the end of a sign and the beginning of the next sign have on each other. In some of the cases, a gesture could be similar to a sub-part of a longer gesture also referred as the *"subgesture problem"* [93]. In natural settings, gestures occur intermittently. Gesture spotting is used to locate the starting point and the endpoint of a gesture in a continuous stream of motion. Once gesture boundaries have been determined, the gesture can be extracted and classified. However, gesture spotting remains challenging for the following reasons:

- Gesture boundaries vary from one signer to another.
- Gesture boundaries are influenced by the surrounding sequence of gestures; they occur in the context of an unfolding interaction [94], and are recognized by higher-level cognitive processes.
- Gestures are unconstrained and cannot be enumerated exhaustively. Because the human body can assume virtually an infinite number of poses during the performance of a gesture, a generic model-based approach to gesture segmentation is not viable.

Gesture spotting is further complicated by individual differences among signers in the shape and duration of the same gesture and variations in how the same signer makes a gesture at different times. An ideal gesture recognizer should be able to extract gesture segments from a video stream and match them against reference patterns

or templates robustly despite high spatio-temporal variability. Lee *et al.* addressed this problem by proposing a model to calculate the likelihood threshold of an input pattern [95]. Fang *et al.* proposed a set of transition movement models (TMMs) for transitions between adjacent signs [96]. Song *et al.* used a latent dynamic conditional random field (LDCRF) with a temporal sliding window to perform online labelling and segmentation of the input sequence [97]. Bhuyan *et al.* proposed a scheme for recognizing different kinds of continuous gestures [98].

3.2.2 Feature extraction: After gesture modelling, a set of features needs to be extracted for gesture recognition. Features for static gestures are derived from image information like colour and texture or pose information like orientation, shape, etc. For dynamic gestures, features are based on motion and/or deformation information like position, skewness, and the velocity of hands. Samples of dynamic hand gestures are spatio-temporal patterns. A static hand gesture may be viewed as a special case of a dynamic gesture with no temporal variation of the hand shape and position. A gesture model should consider both the spatial and temporal characteristics of the hand and its movements. No two samples of the same gesture will result in exactly the same hand and arm motions or the same set of visual images *i.e.*, gestures suffer from spatio-temporal variation. There exists spatio-temporal variation when a user performs the same gesture at different times. Every time the user performs a gesture the shape and the speed of the gesture generally vary. Even if the same person tries to perform the same sign twice, small variation in speed and position of the hands will occur. Therefore, extracted features should be rotation-scaling-translation (RST) invariant. Various features or descriptors are used in the state-of-the-art methods for VGR systems. These features can be broadly classified based on their method of extraction, such as spatial domain features, transform domain features, curve fitting-based features, histogram-based descriptors, and interest point-based descriptors. Moreover, the classifier can handle spatio-temporal variations. Recently, feature extraction techniques based on deep learning have often been applied to gesture recognition. Kong *et al.* [99] proposed a view-invariant feature extraction method using deep learning for multiview action recognition. Table 2 gives a brief survey on the properties of different features used for both static and dynamic gesture recognition.

3.3 Recognition

The final stage of a VGR system is the recognition stage where a suitable classifier recognizes the incoming gesture parameters or features and groups them into predefined classes (supervised) or by their similarity (unsupervised). There are various classifiers used for both static and/or dynamic gesture recognition, each with its own limitations. The most frequently used pattern recognition algorithms applicable to static and/or dynamic gesture recognition are described below.

3.3.1 k means: This algorithm [111, 112] is an unsupervised classifier that determines k centre points to minimize clustering error, defined as the sum of the distances of all data points to their respective cluster centres. The classifier randomly locates k cluster centres in the feature space. Each point in the input dataset is assigned to the nearest cluster centre, and their locations are updated to the average location value for each cluster. This process is then repeated until a stopping condition is met. The stopping condition could be either a user specified maximum number of iterations or a distance threshold for the movement of cluster centres. Ghosh and Ari [113] used a k means clustering based radial basis function neural network (RBFNN) for static hand gesture recognition. In this work, k means clustering is used to determine the RBFNN centres.

3.3.2 Mean shift clustering: Mean shift is a nonparametric clustering method that does not require prior knowledge of the number or distribution of the clusters [114]. The points in the d-dimensional feature space are treated as an empirical probability density function with dense regions at the local maxima or modes of the underlying distribution. A gradient ascent procedure is performed on the local estimated density for each point in the feature space until convergence. The stationary points obtained in this process serve as the modes of the distribution. The data points associated with the same stationary point belong to the same cluster. Bradski proposed a modified mean shift algorithm, continuously adaptive mean shift (CamShift) [115]. CamShift was intended for head and face tracking but can also be used for hand gesture recognition [116].

3.3.3 k-nearest neighbor (kNN): kNN is a non-parametric, supervised learning algorithm. Data in the feature space can be multidimensional. The training data consists of a set of labelled feature vectors. The number k determines how many neighbors (nearby feature vectors) influence the classification [117]. Typically, an odd value of k is chosen for a 2-class classification. Each neighbour may be given the same weight or those closer to the input data may be given more weight (e.g., by applying a Gaussian function). In uniform voting a new feature vector is assigned to the class to which the plurality of its neighbours belong. An alternative method is given in [118]. Hall *et al.* assumed two statistical models (Poisson and binomial) for the sample data to obtain the optimum value of k [119]. The kNN can be used in different applications such as hand gesture-based control media player control [120], sign language recognition [121], etc.

3.3.4 Support vector machine (SVM): An SVM is a supervised classifier for both linearly separable and nonseparable data [122]. This method non-linearly maps the input data (if not linearly separable in current feature space) to some higher dimensional space, where the data can be linearly separated. This mapping from lower to higher dimensional spaces makes the classification of the input data simpler and more accurate. SVMs are often used for hand gesture recognition [123–126]. SVMs were originally designed for two-class classification, and an extension for multi-class classification is necessary for gesture recognition. Weston and Watkins [127] proposed an SVM structure to solve multi-class pattern recognition problem using single optimization stage. Dardas *et al.* [108] used this method along with bag-of-features for hand gesture recognition. However, their single optimization procedure found out to be very complicated to be implemented for real life pattern classification problems [128]. Instead of using single optimization stage, multiple binary classifiers can be used to solve multi-class classification problem, such as "one-against-all" and "one-against-one" methods. Murugeswari and Veluchamy [129] used "one-against-one" multi-class SVM for gesture recognition. It was found that "one-against-one" method performs better than rest of the methods [128].

3.3.5 Dynamic time warping (DTW): DTW can find the optimal alignment of two signals in the time domain. Each element in a time series is represented by a feature vector. Hence, the DTW algorithm calculates the distance between each possible pair of points in two time series in terms of their feature vectors. DTW has been used for gesture recognition by several authors [130, 131].

3.3.6 Finite state machine (FSM): A gesture can be modelled as an ordered state sequence in a spatio-temporal domain using an FSM [132, 133]. The number of states varies by application. The gesture is represented by the trajectory of the hand in 2D space. Offline training is performed using a set of training data for each gesture to derive the FSM parameters for each state. The recognition of gestures can be performed online using the trained FSM. Based on the input value, the recognizer determines whether to make a state transition or to remain in the same state. A gesture has been recognized if the recognizer reaches the final state. More than one model may reach the final state for the same input data. In that case, winning criteria can be applied to choose the most probable gesture.

3.3.7 Hidden Markov model (HMM): An HMM is a doubly stochastic process comprised of 1) an underlying unobservable finite state Markov process and 2) a set of random functions, each associated with a state, that produce an observable output at discrete intervals [134]. The states in the hidden stochastic layer are governed by a set of probabilities:

Table 2 Features used for gesture recognition

Feature type	Examples	Static	Dynamic	Advantages	Limitations
Spatial domain (2D)	Fingertips location, finger direction, and silhouette [100]	✓		• Easy to extract • Rotation invariant.	• Unreliable under occlusion or varying illumination. • Object view-dependent. • Distorted hand trajectory distorts MCC also.
	Motion chain code (MCC) [95, 101]		✓		
Spatial domain (3D)	Joint angles, hand location, surface texture and surface illumination [80]	✓		• 3D modelling can most accurately represent the state of a hand, and thus can give higher recognition accuracy.	• Difficult to accurately estimate 3D shape information of a hand.
Transform domain	Fourier descriptor [102], DCT descriptor [103], Wavelet descriptor [104]	✓	✓	• RST invariant	• Not able to perfectly distinguish different gestures.
Moments	Geometric moments, orthogonal moments [105]	✓	✓	• Moments can be used to derive RST invariant global features.	• Moments are in general global features. So, moments cannot effectively represent an occluded hand.
Curve fitting-based	Curvature Scale Space [106]		✓	• RST invariant. • Resistant to noise.	• Sensitive to distortion in the boundary.
Histogram-based	Histogram of Gradient (HoG) features [107]	✓	✓	• Invariant to geometry and illumination changes.	• Performance is not so satisfactory for images with a complex background and noise.
Interest point-based	SIFT [108], SURF [109]	✓		• RST and illumination invariant	• They are not the best choice for real-time applications because they are computationally expensive.
Mixture of features	Combined features [110]	✓	✓	• Incorporates the advantages of different types of features.	• Classification performance may degrade due to *curse of dimensionality*.

i. The state transition probability distribution $\mathbf{A}$, which gives the probability of transition from the current state to the next possible state.
ii. The observation symbol probability distribution $\mathbf{B}$, which gives the probability of an observation for the present state of the model.
iii. The initial state distribution $\mathbf{\Pi}$, which gives the probability of a state being an initial state.

An HMM is expressed as $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$. HMMs are often used for dynamic gesture recognition [95, 135–137].

The modelling of a gesture sequence involves two phases - feature extraction and HMM training. In the first phase, a particular gesture trajectory is represented by a set of feature vectors. Each of these feature vectors describes dynamics of a hand corresponding to a particular state of a gesture. The number of such states depends on the nature and complexity of a gesture. In the second phase, the vector set is used as an input to HMM. For a given observation sequence, the key issues of HMM are,

- Determination of the probability that the model will generate the observed sequence (forward–backward algorithm).
- Train and adjust the model to maximize the observation sequence probability (Baum–Welch algorithm).
- Determination of the optimal state sequence that produces the observation sequence (Viterbi algorithm).

3.3.8 Conditional random field (CRF): A CRF is a type of discriminative undirected probabilistic graphical model, an alternative to generative HMM model, which predicts a label for each sample of the object for structure prediction considering the context of neighbouring samples. CRFs can be used to enhance recognition accuracy by removing the requirement of conditional independence of observations (as require by generative models like HMM). Bhuyan *et al.* proposed a classification technique based on CRFs on a novel set of motion chain code features [101].

3.3.9 Artificial neural network (ANN): An ANN is a biologically inspired statistical learning algorithm for functional approximation, pattern recognition and classification. ANNs can be used as a supervised classifier for gesture recognition. Training is performed using a set of labelled input patterns. The ANN classifies new input patterns within the labelled classes. ANNs can be used to recognize static [138] and continuous hand gestures and gestures using a 3D articulated hand model [139]. They have been used with a Kinect sensor [140]. A limitation of classical ANN architectures is their inability to handle temporal sequences of features efficiently and effectively [34]. Mainly, they are unable to compensate for changes in temporal shifts and scales especially in real-time applications [42]. Out of several modified architectures, multi-state time-delay neural networks [141] can handle such changes to some extent using dynamic programming. Fuzzy-based neural networks have also been used to recognize gestures [142].

3.3.10 Deep networks: Deep networks are capable of finding salient latent structures within unlabelled and unstructured raw data, and can be used for both feature extraction and classification [143]. Convolutional Neural Networks (CNN) [144, 145], Recurrent Neural Networks (RNN) [146, 147] are such type of deep networks which can be used to handle spatio-temporal variations in gesture

Table 3 Advantages and limitations of different classifiers used in gesture recognition

Classifier	SG[1]	DG[2]	Advantages	Limitations
k-**means** [113]	✓		• Ease of implementation and high-speed performance. • Computationally efficient than hierarchical clustering. • Compact clustering than hierarchical clustering for globular clusters.	• Dependent on initial cluster center values. • Sensitive to outliers. • Does not perform well in the presence of non-globular clusters. • Selecting an appropriate value of k is challenging.
Mean-shift [116]	✓		• No model assumptions unlike GMM or k-means. • Can model non-convex shaped clusters. • Does not depend on initialization unlike k-means. • More resistant to outliers than k-means.	• The algorithm is computationally complex. • Data needs to be sufficiently dense with a discernible gradient to locate the cluster centres. • Susceptible to outliers or data points located between natural clusters.
k-**NN** [120, 121]	✓		• No assumptions about the characteristics of data. • Robust to noisy data. • Ease of implementation.	• Computationally intensive for a large dataset. • Performance degrades as the dimensionality of the feature space increases
SVM [123–126, 129]	✓	✓	• Produces unique solution. • No local minima trapping due to Quadratic Programming (QP). • Presence of regularization reduces over-fitting. • Kernel trick helps to classify non-linearly separable data.	• Right kernel function selection is challenging. • Computationally expensive. • SVM is a binary classifier; and hence, multi-class classification requires multiple pairwise classifications. • Multi-class SVMs based on single optimization [127] are difficult to implement.
DTW [130, 131]		✓	• Reliable alignment of two temporal patterns of different lengths. • No full-model retraining for an inclusion/exclusion of a training gesture. • Independent template computation.	• Requires a large number of training sample for the selection of appropriate templates. • DTW has to align the test and prototype trajectories during each classification. This increases the computational load of classification for large gesture vocabulary size.
FSM [132, 133]		✓	• Can handle patterns of different lengths and states. • An FSM can adapt quickly to accommodate spatio-temporal variability during the training phase, and is therefore robust to any variation in length of the input gesture sequences. • Gesture model is available immediately in an FSM based system.	• Large FSM with many states and transitions can be difficult to manage and maintain. • The state transition conditions are rigid.
HMM [95, 135–137]		✓	• Handles spatio-temporal variations of gestures better than FSM.	• The number of states and the structure of the HMM must be predefined. • Statistical nature of an HMM precludes a rapid training phase. Well-aligned data segments are required to train an HMM. • The stationarity assumption in HMM may not hold true for a complete gestural action.
CRF [101]		✓	• No requirement of conditional independence of observations. • Handles label bias problem.	• High computational complexity during training makes it difficult to re-train the model when new training gesture sequences become available. • CRFs can not recognize totally unknown gestures, *i.e.*, gestures which are not present in the training dataset.
ANN [141, 142]	✓	✓	• Non-parametric, non-linear model. • Adaptive to complex/abstract problems.	• Difficult to set parameters (e.g., the optimal number of nodes, hidden layers, sigmoid function). • Training is computationally intensive and requires a large set of training data for obtaining acceptable performance. • It acts like a "black box," and hence, it is difficult to identify errors in a complex network.
Deep networks [144–153]	✓	✓	• Significant reduction in feature engineering process. • Distributed representation of data.	• A potential problem of deep networks is that it is difficult to get an optimized solution for non-convex and nonlinear systems. • Require large amount of training data.

[1]SG = *Static gesture*; [2]DG = *Dynamic gesture*

recognition. The convolution and pooling layers in CNN can capture discriminative features along both spatial and temporal dimensions to deal with the spatio-temporal variations in gestures. However, CNN can only exploit limited local temporal information and hence, the researchers have moved towards RNN, which can process temporal data using recurrent connections in hidden layers [148]. However, the main drawback of RNN is its short-term memory, which is insufficient for real life temporal variations in gestures. To solve this problem, Long Short-Term Memory (LSTM) [149] was proposed which can tackle longer-range temporal variations. LSTM-based deep networks can be used for efficient modelling of gestures [150–152]. The spatio- temporal graphs are well known for modelling of a spatio-temporal structure. Hence, a combination of high-level spatio-temporal graphs and RNN can also be used to solve spatio-temporal modelling problem of RNN [153]. Deep learning techniques can give outstanding performance in both feature extraction and recognition owing to their built-in feature learning capability. The effective and efficient algorithms of deep networks are capable of solving complex optimization tasks.

The main advantages and limitations of different classifiers used for static and/or dynamic gesture recognition are enumerated in Table 3.

4 Conclusion

Hand gesture recognition is an important research area in computer vision with many applications to human–computer interaction. Use of gesture-based interface has a great potential in assisting humans in their day to day problems. In recent years, vision-based gesture recognition is successfully used in different applications such as healthcare and medicine, virtual reality, and driver monitoring. However, vision-based recognition is extremely challenging not only because of its diverse contexts, multiple interpretations, and spatio-temporal variations but, also because of the complex non-rigid properties of the human hand. The existing classifiers used for vision-based gesture recognition are not capable of simultaneously handling all the gesture classification problems. Each has drawbacks limiting overall performance.

Vision-based gesture recognition typically depends on three stages: gesture detection and pre-processing; gesture representation and feature extraction; and recognition. Detection includes various kind of imaging systems whereas pre-processing includes segmentation of gesturing body parts and occlusion handling. Accurate detection of gestures is difficult because of varying illumination, shadows, complex background, and other factors. The complex articulated shape of the hand makes it harder to model the appearance of both static and dynamic gestures. Variation of gesture parameters due to spatio-temporal variance in hand postures makes the recognition process more difficult. A gesture can be represented using an appropriate model based on the spatio-temporal properties of the gesture. Each of these models have their own advantages and disadvantages, which affect the overall performance of a gesture recognition system. Gesture modelling is followed by the selection and extraction of appropriate features. Extracted features should be RST invariant, object view independent, and computationally inexpensive. Selection of a feature also depends on the nature of gestures. The third stage of gesture recognition module consists of a classifier, which classifies the input gesture into an unknown (unsupervised) or a known class (supervised). However, every classifier has its own limitations. This paper surveyed the major constraints in the different stages of a VGR system. A vision-based interface is expected to enable a user to interact with a machine with natural *human-to-human* interactions in an unconstrained environment. Its performance should be accurate irrespective of changes in the environment. However, most existing VGR systems are only work in a limited range of indoor applications. Also, the same gesture can have different meanings in different cultures. Thus, a lack of cross-cultural conventions limits their global acceptance. Feedback is also required to indicate the correctness of the input gestures. Feedback provides the user some means of self-learning of the necessary gestures and the contexts to be used. Present gesture-based interfaces generally lack required feedback for VGR system validation.

5 References

1 'ACM SIGCHI curricula for human–computer interaction', http://www.acm.org/sigchi/cdg/cdg2.html, accessed 1992
2 Preece, J., Carey, T., Rogers, Y., *et al.*: 'Human-computer interaction', (Pearson Educ. Ltd., 1994)
3 Dix, A., Finlay, J. E., Abowd, G. D., *et al.*: 'Human-computer interaction', (Prentice-Hall, Inc., 2003, 3rd ed.)
4 Jaimes, A., Sebe, N.: 'Multimodal human-computer interaction: A survey', Comput. Vis. Image Underst., 2007, **108**, (1–2), pp. 116–134
5 Moeslund, T. B., Hilton, A., Krüger, V.: 'A survey of advances in vision-based human motion capture and analysis', Comput. Vis. Image Underst., 2006, **104**, (2), pp. 90–126
6 Sandbach, G., Zafeiriou, S., Pantic, M., *et al.*: 'Static and dynamic 3D facial expression recognition: A comprehensive survey', Image Vis. Comput., 2012, **30**, (10), pp. 683–697
7 Kumar, S.; Bhuyan, M.K.; Chakraborty, B. K.: 'Extraction of informative regions of a face for facial expression recognition', IET Comp. Vis., 2016, 10, (6), p. 567–576
8 Song, F., Tan, X., Chen, S., *et al.*: 'A literature survey on robust and efficient eye localization in real-life scenarios', Pattern Recogn., 2013, **46**, (12), pp. 3157–3173
9 Mitra, S., Acharya, T.: 'Gesture recognition: A survey', IEEE Trans. Syst., Man, Cybern., Part C: Applicat. and Reviews, 2007, **37**, (3), pp. 311–324
10 El Ayadi, M., Kamel, M. S., Karray, F.: 'Survey on speech emotion recognition: Features, classification schemes, and databases', Pattern Recogn., 2011, **44**, (3), pp. 572–587
11 Kinnunen, T., Li, H.: 'An overview of text-independent speaker recognition: From features to supervectors', Speech Comm., 2010, **52**, (1), pp. 12–40
12 Li, J., Deng, L., Gong, Y., *et al.*: 'An overview of noise-robust automatic speech recognition', IEEE/ACM Trans. Audio, Speech, and Language Process., 2014, **22**, (4), pp. 745–777
13 Kumar, P., Rautaray, S., Agrawal, A.: 'Hand data glove: A new generation real-time mouse for human-computer interaction', Proc. 1st Int. Conf. Recent Advances in Inform. Technol., 2012, pp. 750–755
14 Pantic, M., Rothkrantz, L. J. M.: 'Toward an affect-sensitive multimodal human-computer interaction', Proc. IEEE, 2003, **91**, (9), pp. 1370–1390
15 Oviatt, S.: 'Multimodal interfaces', in The Human–Comput. Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applicat. (Human Factors and Ergonomics Series), A. Sears and J. A. Jacko, Eds. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 2007, ch. 18.
16 Jiang, R., Sadka, A., Crookes, D.: 'Multimodal biometric human recognition for perceptual human-computer interaction', IEEE Trans. Syst., Man, Cybern., Part C: Applicat. and Reviews, 2010, **40**, (6), pp. 676–681
17 Lichtenauer, J. F., Hendriks, E. A., Reinders, M. J. T.: 'Sign Language Recognition by Combining Statistical DTW and Independent Classification', PAMI, 2008, **30**, (11), pp. 2040–2046
18 Sagayam, K. M., Hemanth, D. J.: 'Hand posture and gesture recognition techniques for virtual reality applications: a survey', Virtual Reality, 2017, **21**, (2), pp. 91–107
19 Kulshreshth, A., Pfeil, K., LaViola, J. J.: 'Enhancing the Gaming Experience Using 3D Spatial User Interface Technologies', IEEE Comp. Graphics and Appl., 2017, **38**, (3), pp. 16–23
20 Reifinger, S., Wallhoff, F., AblaÃ§meier, M., *et al.*: 'Static and dynamic hand-gesture recognition for augmented reality applications', Proc. Int. Conf. HumanâĂŞComputer Interaction, Beijing, China, 2007, pp. 728-âĂŞ737
21 McCowan, I., Gatica-Perez, D., Bengio, S., *et al.*: 'Automatic analysis of multimodal group actions in meetings', PAMI, 2005, **27**, (3), pp. 305–317
22 Wei Lun Ng, Chee Kyun Ng, Nor Kamariah Noordin, *et al.*: 'Gesture Based Automating Household Appliances', Proc. Int. Conf. HumanâĂŞComputer Interaction, Beijing, China, 2011, pp. 285-âĂŞ293
23 Jacob, M., Cange, C., Packer, R., *et al.*: 'Intention, context and gesture recognition for sterile MRI navigation in the operating room', Progress in Pattern Recogn., Image Anal., Comput. Vis., and Applicat., ser. Lecture Notes in Comput. Sci., L. Alvarez, M. Mejail, L. Gomez, and J. Jacobo, Eds. Springer Berlin Heidelberg, 2012, **7441**, pp. 220âĂŞ-227
24 Tao, L., Zappella, L., Hager, G., *et al.*: 'Surgical gesture segmentation and recognition', Med. Image Comput. and Comput.- Assisted Intervention MICCAI 2013, ser. Lecture Notes in Comput. Sci., K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds. Springer Berlin Heidelberg, 2013, **8151**, pp. 339âĂŞ-346.
25 Kim-Tien, N., Truong-Thinh, N., Cuong, T.: 'A method for controlling wheelchair using hand gesture recognition', Robot Intell. Technol. and Applicat. 2012, ser. Advances in Intell. Syst. and Comput., J.-H. Kim, E. T. Matson, H. Myung, and P. Xu, Eds. Springer Berlin Heidelberg, 2013, **208**, pp. 961-âĂŞ970
26 Smith, P., Shah, M., da Vitoria Lobo, N.: 'Determining driver visual attention with one camera', IEEE Trans. Intell. Transp. Syst., 2003, **4**, (4), pp. 205–218
27 Pickering, C.: 'The search for a safer driver interface: a review of gesture recognition human machine interface', J. Comput. Control Eng., 2005, **16**, (1), pp. 34âĂŞ-40
28 Zeng, B., Wang, G., Lin, X.: 'A hand gesture based interactive presentation system utilizing heterogeneous cameras', Tsinghua Sci. and Technol., 2012, **17**, (3), pp. 329-âĂŞ336
29 Hariharan, B., Padmini, S., Gopalakrishnan, U.: 'Gesture recognition using Kinect in a virtual classroom environment', Proc. Fourth Int. Conf. Digital Info.

Comm. Tech. and its App. (DICTAP), Bangkok, 2014, pp. 118–124

30 Arafa, Y., Mamdani, A.: 'Building multi-modal personal sales agents as interfaces to e-commerce applications', in Active Media Technol., in Lecture Notes in Comput. Sci., J. Liu, P. Yuen, C.-h. Li, J. Ng, and T. Ishida, (Eds. Springer Berlin Heidelberg, 2001, **2252**, pp. 113–133)

31 Zhang, Z.: 'Microsoft Kinect Sensor and Its Effect', in IEEE MultiMedia, 2012, **19**, (2), pp. 4–10

32 Kumara, W. G. C. W., Wattanachote, K., Battulga, B., *et al.*: 'A Kinect-based assessment system for smart classroom', Int. J. Dist. Edu. Tech. (IJDET), 2017, **13**, (2), 34–53

33 'SoftKinetic's gesture control technology rolls out in additional car model', https://www.softkinetic.com/AboutUs/NewsEvents/ArticleView/ArticleId/545/PRESS-RELEASE-SoftKinetic-s-Gesture-Control-Technology-Rolls-Out-in-Additional-Car-Model.html, accessed May 2017.

34 Pavlovic, V., Sharma, R., Huang, T.: 'Visual interpretation of hand gestures for human-computer interaction: a review', PAMI, 1997, **19**, (7), pp. 677–695

35 Wu, Y., Huang, T. S.: 'Vision-based gesture recognition: A review', Proc. Int. Gesture Workshop on Gesture-Based Commun. in Human-Comput. Interaction, 1999, pp. 103–115

36 Moeslund, T. B., Granum, E.: 'A survey of computer vision-based human motion capture', Comput. Vis. and Image Underst., 2001, **81**, (3), pp. 231–268

37 Derpanis, K. G.: 'A review of vision-based hand gestures', Dept. of Comput. Sci. York University, Internal Rep., 2004

38 Erol, A., Bebis, G., Nicolescu, M., *et al.*: 'Vision-based hand pose estimation: A review', Comput. Vis. Image Underst., 2007, **108**, (1–2), pp. 52–73

39 Moni, M. A., Ali, A.: 'Hmm based hand gesture recognition: A review on techniques and approaches', Proc. 2nd IEEE Int. Conf. Comput. Sci. and Inform. Technol. (ICCSIT), Aug. 2009, pp. 433–437

40 Wachs, J. P., Kölsch, M., Stern, H., *et al.*: 'Vision-based hand-gesture applications', Commun. ACM, 2011, **54**, (2), pp. 60–71

41 Suarez, J., Murphy, R.: 'Hand gesture recognition with depth images: A review', Proc. IEEE RO-MAN, 2012, pp. 411–417

42 Rautaray, S., Agrawal, A.: 'Vision based hand gesture recognition for human computer interaction: a survey', Artificial Intell. Review, 2012, **43**, (1), pp. 1–54

43 Hasan, H., Abdul-Kareem, S.: 'Human-computer interaction using vision-based hand gesture recognition systems: a survey', Neural Comput. and Applicat., 2011, pp. 1–11

44 Cheng, H., Yang, L., Liu, Z.: 'Survey on 3D hand gesture recognition', IEEE Trans. Circuits Syst. Video Technol., 2016, **26**, (9), pp. 1659–1673

45 Karam, M.: 'A framework for research and design of gesture based human-computer interactions', Ph.D. dissertation, University of Southampton, Oct. 2006.

46 Campbell, L., Becker, D., Azarbayejani, A., *et al.*: 'Invariant features for 3-d gesture recognition', Proc. 2nd Int. Conf. Automat. Face and Gesture Recogn., Oct. 1996, pp. 157–162

47 Bhuyan, M. K., Ghosh, D., Bora, P.: 'Feature extraction from 2D gesture trajectory in dynamic hand gesture recognition', Proc. IEEE Conf. Cybern. and Intell. Syst., Jun. 2006, pp. 1–6

48 Bhuyan, M. K., Ghosh, D., Bora, P.: 'Continuous hand gesture segmentation and co-articulation detection', Computer Vision, Graphics and Image Processing, in Lecture Notes in Computer Science, P. Kalra and S. Peleg, (Eds. Springer Berlin Heidelberg, 2006, **4338**, pp. 564–575)

49 Bhuyan, M. K., Ghosh, D., Bora, P.: 'Estimation of 2D motion trajectories from video object planes and its application in hand gesture recognition', Pattern Recognition and Machine Intelligence, in Lecture Notes in Computer Science, S. Pal, S. Bandyopadhyay, and S. Biswas, (Eds. Springer Berlin Heidelberg, 2005, **3776**, pp. 509–514)

50 Berman, S., Stern, H.: 'Sensors for gesture recognition systems', IEEE Trans. Syst., Man, Cybern.,Syst. C., Appl., Rev., 2012, **42**, (3), pp. 277–290

51 Socolinsky, D. A., Selinger, A.: 'A comparative analysis of face recognition performance with visible and thermal infrared imagery', Proc. Int. Conf. Pattern Recog., 2002, **4**, pp. 217–222

52 Zhenyao, M. and Neumann, U.: 'Real-time hand pose recognition using low-resolution depth images', Proc. CVPR, 2006, pp. 1499–1505

53 Wang, Y., Yu, T., Shi, L., *et al.*: 'Using human body gestures as inputs for gaming via depth analysis', Proc. IEEE Int. Conf. Multimedia and Expo, Jun. 2008, pp. 993–996

54 Cerlinca, T., Pentiuc, S.: 'Robust 3D hand detection for gestures recognition', in Intell. Distrib. Comput. V, ser. Stud. in Comput. Intell., F. Brazier, K. Nieuwenhuis, G. Pavlin, M. Warnier, and C. Badica, (Eds. Springer Berlin Heidelberg, 2012, **382**, pp. 259–264)

55 Hongo, H., Ohya, M., Yasumoto, M., *et al.*: 'Focus of attention for face and hand gesture recognition using multiple cameras', Proc. Fourth IEEE Int. Conf. . Auto. Face and Gesture Recogn., 2000, pp. 156–161

56 Tsai, C. Y., Lee, Y. h.: 'Multiple-camera-based gesture recognition by MDA method', Proc. Fifth Int. Conf. Fuzzy Sys. and Knowledge Discovery, 2008, **3**, pp. 599–603

57 Ogawara, K., Takamatsu, J., Hashimoto, K., *et al.*: 'Grasp recognition using a 3D articulated model and infrared images', Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS), Oct. 2003, pp. 1590–1595

58 Droeschel, D., Stuckler, J., Behnke, S.: 'Learning to interpret pointing gestures with a time-of-flight camera', Proc. 6th ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), Mar. 2011, pp. 481–488

59 Uebersax, D., Gall, J., Van den Bergh, M., *et al.*: 'Real-time sign language letter and word recognition from depth data', Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops), Nov. 2011, pp. 383–390

60 Gonzalez-Sanchez, T., Puig, D.: 'Real-time body gesture recognition using depth camera', Electron. Lett., 2011, **47**, (12), pp. 697–698

61 'Introduction to the Time-of-Flight (ToF) system design - user's guide', http://www.ti.com/lit/ml/sbau219d/sbau219d.pdf, accessed 2014.

62 'Basler ToF Camera. USER'S MANUAL', https://www.baslerweb.com/en/support/ downloads/document-downloads/basler-tof-camera-users-manual/.html, accessed 2016.

63 Ren, Z., Yuan, J., Meng, J., *et al.*: 'Robust part-based hand gesture recognition using Kinect sensor', IEEE Trans. Multimedia, 2013, **15**, (5), pp. 1110–1120

64 Rioux-Maldague, L., Giguere, P.: 'Sign language finger spelling classification from depth and color images using a deep belief network', Proc. Canadian Conf. Comput. and Robot Vis. (CRV), May 2014, pp. 92–97

65 Regenbrecht, J. C. H., Hoermann, S.: 'A leap-supported, hybrid interface approach', Proc. Austral. Comput.-Human Interact. Conf., 2013, pp. 281–284

66 Lu, W., Tong, Z., Chu, J.: 'Dynamic hand gesture recognition with leap motion controller', IEEE Signal Proc. Lett., 2016, **23**, (9), pp. 1188–1192

67 Porfirio, L. E. S. O. A. J., Lais Wiggers, K., Weingaertner, D.: 'Libras sign language hand configuration recognition based on 3D meshes', Proc. IEEE SMC, Oct. 2013, p. 1588–1593.

68 Juang, C.-F., Chiu, S.-H., Shiu, S.-J.: 'Fuzzy system learned through fuzzy clustering and support vector machine for human skin color segmentation', IEEE Trans. Syst., Man, Cybern., Part A: Syst. and Humans, 2007, **37**, (6), pp. 1077–1087

69 Juang, C.-F., Chang, C.-M., Wu, J.-R., *et al.*: 'Computer vision based human body segmentation and posture estimation', IEEE Trans. Syst., Man, Cybern., Part A: Syst. and Humans, 2009, **39**, (1), pp.119–133

70 Powar, V., Jahagirdar, A., Sirsikar, S.: 'Skin detection in YCbCr color space', IJCA Proc. Int. Conf. Comput. Intell. Mar. 2012

71 Störring, M., Andersen, H. J., Granum, E.: 'Skin color detection under changing lighting conditions', Proc. 7th Symp. Intell. Robot. Syst., 1999, pp. 187–195

72 Chakraborty, B. K., Bhuyan, M. K., Kumar, S.: 'Fusion-based Skin Detection Using Image Distribution Model', Proc. Tenth Indian Conf. Comp. Vis., Graphics and Image Process., 2016, pp. 67:1–67:8

73 Chakraborty, B. K., Bhuyan, M. K., Kumar, S.: 'Combining image and global pixel distribution model for skin colour segmentation', Pattern Recognition Letters, 2017, **41**, pp. 33–40

74 Kawulok, M., Kawulok, J., Nalepa, J.: 'Spatial-based skin detection using discriminative skin-presence features', Pattern Recognition Letters, 2014, **8**, pp. 3–13

75 Chakraborty, B. K., Bhuyan, M. K.: 'Skin segmentation using possibilistic fuzzy c-means clustering in presence of skin-colored background', Proc. IEEE Recent Advances in Intel. Comp. Syst., Dec. 2015, pp. 246–250

76 Lee, J., Kunii, T.: 'Constraint-based hand animation', Models and Techn. in Comput. Animation, ser. Comput. Animation Series, N. Thalmann and D. Thalmann, (Eds. Springer Japan, 1993, pp. 110–127)

77 Shotton, J., Fitzgibbon, A., Cook, M., *et al.*: 'Real-time human pose recognition in parts from single depth images', Proc. CVPR, June 2011, pp. 1297–1304

78 Keskin, C., Kirac, F., Kara, Y. E., *et al.*: 'Real time hand pose estimation using depth sensors', Springer, 2013, pp. 119–137

79 Yao, Y., Fu, Y.: 'Real-time hand pose estimation from RGB-D sensor', Proc. IEEE Int. Conf. Multi. and Expo, July 2012, pp. 705–710

80 de La Gorce, M., Paragios, N.: 'A variational approach to monocular hand-pose estimation', Comput. Vis. Image Understand., 2010, **114**, (3), pp. 363–372

81 de La Gorce, M., Fleet, D. J., Paragios, N.: 'Model-based 3D hand pose estimation from monocular video', PAMI, 2011, **33**, (9), pp. 1793–1805

82 Oikonomidis, I., Kyriazis, N., Argyros, A.: 'Markerless and efficient 26-DOF hand pose recovery', Proc. ACCV, 2010, pp. 744–757

83 Iason Oikonomidis, N. K., Argyros, A.: 'Efficient model-based 3D tracking of hand articulations using kinect', Proc. BMVC, 2011, pp. 101.1–101.11

84 Wang, J., Liu, Z., Chorowski, J., *et al.*: 'Robust 3D action recognition with random occupancy patterns', Proc. ECCV, 2012, pp. 872–885

85 Liang, B.: 'Gesture recognition using depth images', Proc. 15th ACM ICMI, 2013, pp. 353–356

86 Ohn-Bar, E., Trivedi, M. M.: 'The power is in your hands: 3D analysis of hand gestures in naturalistic video', Proc. IEEE Conf. CVPR Workshops, June 2013, pp. 912–917

87 Sohn, M.-K., Lee, S.-H., Kim, D.-J., *et al.*: 'A comparison of 3D hand gesture recognition using dynamic time warping', Proc. 27th Conf. Image Vis. Comput., 2012, pp. 418–422

88 Doliotis, P., Stefan, A., McMurrough, C., *et al.*: 'Comparing gesture recognition accuracy using color and depth information', Proc. 4th Int. Conf. Pervasive Technol. Rel. Assistive Environ., 2011, pp. 20:1–20:7

89 Zhu, H. M., Pun, C. M., 'Real-time hand gesture recognition from depth image sequences', Proc. 9th Int. Conf. Comput. Graph., Imag., Visualizat., July 2012, pp. 49–52

90 Xu, D., Chen, Y. L., Lin, C., *et al.*: 'Real-time dynamic gesture recognition system based on depth perception for robot navigation', Proc. IEEE Int. Conf. Robot. and Biomimetics (ROBIO), Dec 2012, pp. 689–694

91 Dondi, P., Lombardi, L., Porta, M.: 'Development of gesture-based human computer interaction applications by fusion of depth and color video streams', IET Comp. Vis., 2014, **8**, (6), pp. 568–578

92 Chai, Y., Shin, S., Chang, K., *et al.*: 'Real-time user interface using particle filter with integral histogram', IEEE Trans. Consumer Elect., 2010, **56**, (2), pp. 510–515

93 Alon, J., Athitsos, V., Yuan, Q. *et al.*: 'A unified framework for gesture recognition and spatiotemporal gesture segmentation', PAMI, 2009, **31**, (9), pp. 1685–1699

94 Dourish, P.: 'What we talk about when we talk about context', Personal Ubiquitous Comput., 2004, **8**, (1), pp. 19–30

95 Lee, H.-K., Kim, J.: 'An HMM-based threshold model approach for gesture recognition', PAMI, 1999, **21**, (10), pp. 961–973

96 Fang, G., Gao, W., Zhao, D.: 'Large-vocabulary continuous sign language recognition based on transition-movement models', IEEE Trans. Syst., Man, Cybern., Part A: Syst. and Humans, 2007, **37**, (1), pp. 1–9
97 Song, Y., Demirdjian, D., Davis, R.: 'Continuous body and hand gesture recognition for natural human-computer interaction', ACM Trans. Interact. Intell. Syst., 2012, **2**, (1), pp. 5:1–5:28
98 Bhuyan, M. K., Bora, P. K., Ghosh, D.: 'An integrated approach to the recognition of a wide class of continuous hand gestures', Int. J. Pattern Recogn. and Artificial Intel., 2011, **25**, (2), pp. 227–252
99 Kong, Y., Ding, Li, J. *et al.*: 'Deeply Learned View-Invariant Features for Cross-View Action Recognition', IEEE Trans. Image Process., 2017, **26**, (6), pp. 3028–3037
100 Nguyen, T. N., Vo, D. H., Huynh, H. H., *et al.*: 'Geometry based static hand gesture recognition using support vector machine', Proc. 13th Int. Conf. Control Auto. Robot. Vis., Dec 2014, pp. 769–774
101 Bhuyan, M. K., Ajay Kumar, D., MacDorman, K., *et al.*: 'A novel set of features for continuous hand gesture recognition', J. Multimodal User Interfaces, 2014, **8**, (4), pp. 333–343
102 Harding, P.R., Ellis, T.J.: 'Recognizing hand gesture using fourier descriptors', Proc. ICPR, 2004, pp. 286–289
103 Akhter, I., Sheikh, Y., Khan, S., *et al.* : 'Trajectory space: A dual representation for nonrigid structure from motion', PAMI, 2011, **33**, (7), pp. 1442–1456
104 Hung, K.C.: 'The generalized uniqueness wavelet descriptor for planar closed curves',IEEE Trans. Image Process, 2000, **9**, (5), pp. 834–845
105 Priyal, S. P., Bora, P. K.: 'A study on static hand gesture recognition using moments', Proc. 2010 Int. Conf. Signal Process. and Comm., July 2010, pp. 1–5
106 Wu, X., Cia, M., Chen, L.,*et al.*: 'Point context: an effective shape descriptor for RST-invariant trajectory recognition', J. Math. Imag. and Vis., 2016, **56**, (3), pp. 441–454
107 Feng, K. P., Yuan, F.: 'Static hand gesture recognition based on hog characters and support vector machines', Proc. 2nd Int. Symp. Inst. and Measure., Sensor Network and Auto., Dec 2013, pp. 936–938
108 Dardas, N., Chen, Q., Georganas, N. D., Petriu, E.M.: 'Hand gesture recognition using Bag-of-features and multi-class Support Vector Machine', Proc. Int. Symp. Haptic Audio Visual Env. and Games, Oct 2010, pp. 1–5
109 Zhang, R., Ming, Y., Sun, J.: 'Hand gesture recognition with SURF-BOF based on gray threshold segmentation', Proc. Int. Signal Process., Nov 2016, pp. 118–122
110 Ghosh, D. K., Ari, S.: 'Static hand gesture recognition using mixture of features and SVM classifier', Proc. 5th Int. Conf. Comm. Sys. and Network Tech., April 2015, pp. 1094–1099.
111 Forgy, E. W.: 'Cluster analysis of multivariate data: efficiency vs interpretability of classifications', Biometrics, 1965, **21**, pp. 768–769
112 Macqueen, J.: 'Some methods for classification and analysis of multivariate observations', Proc. 5th Berkeley Symp. Math. Stat. and Probability, 1967, pp. 281–297
113 Ghosh, D. K., Ari, S.: 'A static hand gesture recognition algorithm using k-mean based radial basis function neural network', Proc. 8th Int. Conf. Info. Comm. Signal Process., 2011, pp. 1–5
114 'Mean shift clustering, lecture notes', http://www.cse.yorku.ca/kosta/CompVis Notes/mean shift.pdf/
115 Bradski, G.: 'Real time face and object tracking as a component of a perceptual user interface', Proc., 4th IEEE Workshop Applicat. of Comput. Vis. (WACV), Oct. 1998, pp. 214–219
116 Nadgeri, S., Sawarkar, S., Gawande, A.: 'Hand gesture recognition using Camshift algorithm', Proc. 3rd Int. Conf. Emerging Trends in Eng. and Technol. (ICETET), Nov. 2010, pp. 37–41
117 'A detailed introduction to knearest neighbor (knn) algorithm', http://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/
118 Coomans, D., Massart, D.: 'Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-nearest neighbour classification by using alternative voting rules', Analytica Chimica Acta, 1982, **136**, pp. 15–27
119 Hall, P., Park, B. U., Samworth, R. J.: 'Choice of neighbor order in nearest-neighbor classification', Ann. Statist., 2008, **36**, (5), pp. 2135–2152
120 Marasovic, T., Papic, V.: 'Feature weighted nearest neighbour classification for accelerometer-based gesture recognition', Proc. 20th Int. Conf. Softw., Telecommun. and Comput. Networks (SoftCOM), Sept. 2012, pp. 1–5
121 Gupta, B., Shukla, P., Mittal, A.: 'K-nearest correlated neighbor classification for Indian sign language gesture recognition using feature fusion', Proc. Int. Conf. Comp. Comm. Info. (ICCCI), Coimbatore, 2016, pp. 1–5.
122 Burges, C. J. C.: 'A tutorial on support vector machines for pattern recognition', Data Min. Knowl. Discov., 1998, **2**, (2), pp. 121–167
123 Dardas, N., Georganas, N. D.: 'Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques', IEEE Trans. Instrum. Meas., 2011, **60**, (11), pp. 3592–3607
124 Keskin, C., Kirac, F., Kara, Y.: 'Real time hand pose estimation using depth sensors', Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops), Nov. 2011, pp. 1228–1234
125 Liu, L., Xing, J., Ai, H., *et al.*: 'Hand posture recognition using finger geometric feature', Proc. 21st Int. Conf. Pattern Recogn. (ICPR), Nov. 2012, pp. 565–568
126 Rodriguez, K. O., Chavez, G. C.: 'Finger spelling recognition from RGB-D information using kernel descriptor', 26th SIBGRAPI Conf. Graph., Patterns and Images (SIBGRAPI), Aug. 2013, pp. 1–7
127 Weston, J., Watkins, C.: 'Multi-class support vector machines', Proc. ESANN, 1999, pp. 219–224
128 Hsu, C.-W., Lin, C.-J.: 'A comparison of methods for multiclass support vector machines', IEEE Trans. Neural Netw., 2002, **13**, (2), pp. 415–425
129 Murugeswari, M., Veluchamy, S.: 'Hand gesture recognition system for real-time application', Proc. Int. Conf. Adv. Comm., Cont. and Comp. Tech., 2014, pp. 1220–1225
130 Wenjun, T., Chengdong, W., Shuying, Z., *et al.*: 'Dynamic hand gesture recognition using motion trajectories and key frames', Proc. 2nd Int. Conf. Advanced Comput. Control (ICACC), vol. 3, Mar. 2010, pp. 163–167
131 Hussain, S., Rashid, A.: 'User independent hand gesture recognition by accelerated DTW', Proc. Int. Conf. Inform., Electron. Vis. (ICIEV), May 2012, pp. 1033–1037
132 Hong, P., Turk, M., Huang, T.: 'Gesture modelling and recognition using finite state machines', Proc. 4th IEEE Int. Conf. Automat. Face and Gesture Recogn., 2000, pp. 410–415
133 Bhuyan, M. K.: 'FSM-based recognition of dynamic hand gestures via gesture summarization using key video object planes', World Academy of Science, Engineering and Technology., 2012, **6**, (68), pp. 724–735
134 Rabiner, L., Juang, B.-H.: 'An introduction to hidden Markov models', IEEE ASSP Mag., 1986, **3**, (1), pp. 4–16
135 Kwon, J., Park, F.: 'Natural movement generation using hidden Markov models and principal components', IEEE Trans. Syst., Man, Cybern., Part B: Syst. and Cybern., 2008, **38**, (5), pp. 1184–1194
136 Heracleous, P., Aboutabit, N., Beautemps, D.: 'Lip shape and hand position fusion for automatic vowel recognition in cued speech for french', IEEE Signal Process. Lett., 2009, **16**, (5), pp. 339–342
137 Peng, S.-Y., Wattanachote, K., Lin, H.-J., *et al.*: 'A real-time hand gesture recognition system for daily information retrieval from internet', Proc. 4th Int. Conf. Ubi-Media Comput., Jul. 2011, pp. 146–151
138 Hasan, H., Abdul-Kareem, S.: 'Static hand gesture recognition using neural networks', Artificial Intell. Review, 2014, **41**, (2), pp. 147–181
139 Nolker, C., Ritter, H.: 'Visual recognition of continuous hand postures', IEEE Trans. Neural Netw., 2002, **13**, (4), pp. 983–994
140 Patsadu, O., Nukoolkit, C., Watanapa, B.: 'Human gesture recognition using Kinect camera', Proc. Int. Joint Conf. Comput. Sci. and Softw. Eng. (JCSSE), May 2012, pp. 28–32
141 Yang, M.-H., Ahuja, N., Tabb, M.: 'Extraction of 2D motion trajectories and its application to hand gesture recognition', PAMI, 2002, **24**, (8), pp. 1061–1074
142 Varkonyi-Koczy, A., Tusor, B.: 'Human-computer interaction for smart environment applications using fuzzy hand posture and gesture models', IEEE Trans. Instrum. Meas., 2011, **60**, (5), pp. 1505–1514
143 Krizhevsky, A., Sutskever, I., Hinton, G. E.: 'ImageNet classification with deep convolutional neural networks', Proc. conf. Advances in Neural Information Processing Systems, 2012, pp. 1106–1114
144 Kim, Y., Toomajian, B.: 'Hand gesture recognition using micro-Doppler signatures with convolutional neural network', IEEE Access, 2016, **4**, pp. 7125–7130
145 D. Wu et al.: 'Deep dynamic neural networks for multimodal gesture segmentation and recognition', PAMI, 2016, **38**, (8), pp. 1583–1597
146 Chai, X., Liu, Z., Yin, F., *et al.*: 'Two streams Recurrent Neural Networks for Large-Scale Continuous Gesture Recognition', Proc. conf. ICPR, 2016, pp. 31–36
147 Du, Y., Wang, W., Wang, L.: 'Hierarchical recurrent neural network for skeleton based action recognition', Proc. conf. CVPR, 2015, pp. 1110–1118
148 Asadi-Aghbolaghi, M., *et al.*: 'A survey on deep learning based approaches for action and gesture recognition in image sequences', Proc. conf. Automatic Face & Gesture Recognition, 2017, pp. 476–483
149 Gers, F. A., Schraudolph, N. N. and Schmidhuber, J.: 'Learning precise timing with LSTM recurrent networks', J. Mach. Learn. Res., 2003, **3**, pp. 115–143
150 Donahue, J., *et al.*: 'Long-term recurrent convolutional networks for visual recognition and description', PAMI, 2017, **39**, (4), pp. 677–691
151 Tsironi, E., Barros, P., Wermter, S.: 'Gesture recognition with a convolutional long short-term memory recurrent neural network', Proc. European Symp. Artificial Neural Net., Comp. Intel. and Machine Learn. (ESANN), 2016, pp. 213–218
152 Liu, J., Shahroudy, A., XuGang Wang, D.: 'Spatio-temporal LSTM with trust gates for 3D human action recognition', Proc. ECCV, 2016, pp. 816–833
153 Jain, A., Zamir, A. R., Savarese, S.: 'Structural-RNN: Deep learning on spatio-temporal graphs', Proc. conf. CVPR, 2016, pp. 5308–5317