Chapter 1

Vision Sensors for Robotic Perception

Shan Luo[1] Daniel Fernandes Gomes[2] Jiaqi Jiang[2] and Guanqun Cao[2]

In this chapter, we will introduce the vision sensors for robotics applications. It first briefly introduces the working principles of the widely used vision sensors, i.e., RGB cameras, stereo cameras and depth sensors, and also the off-the-shelf vision sensors that have been widely used in the robotics research, particularly robot perception. As one of the most widely used sensors to be equipped with robots and thanks to its low cost and high resolution, vision sensors have also been used in other sensing modalities. In recent years, there is a rapid development of embedding vision sensors in optical tactile sensors. In such sensors, visual cameras are placed under an elastomer layer and used to capture its deformation while being interacted with objects. The vision sensors enable robots to sense and estimate the properties of the objects, e.g., object shapes, appearances, textures and mechanical parameters. We will cover various aspects of vision sensors for robotic applications including the various technologies, hardware, integration, computation algorithms and applications in relation to robotics.

1.1 Introduction

Our eyes are crucial for seeing the world around us. With eyes, we can see the color of a cat, texture of a carpet, the face of a person and appearance of a building. Similarly, vision sensors have also been developed in the past decades and can be equipped to robots to enable them to have the sense of sight. Vision sensors provide robots vital information about the surroundings and vision has been the sensing modality robots rely on most.

Compared to other sensing modalities like touch sensing, hearing, smell and taste, vision has a much larger Field of View (FoV) and is able to capture the view of a scene at a glance and multiple objects can be observed in a single view. The properties of objects in the scene, e.g., color, textures, appearances and shapes can be obtained from one single camera image and it is remarkably with ease to collect data with vision sensors..

[1] Department of Engineering, King's College London, London WC2R 2LS, U.K. Email: shan.luo@kcl.ac.uk
[2] smARTLab, Department of Computer Science, University of Liverpool, Liverpool L69 3BX, U.K. Emails: {daniel.fernandes-gomes, jiaqi.jiang, psgcao}@liverpool.ac.uk

On the other hand, processing visual data requires high computational resources. More than 50 percent of the human brain [1] is devoted directly or indirectly to processing visual information and therefore visual information has been a key for humans to understand the world. For robots, much of the processing power is also devoted to extracting information from the visual data. One reason is that the abundant data can be accessed via the vision sensors. The other reason is that there are fluctuation factors in visual data that affect extracting useful information from visual data.

Such factors include scaling, rotation, translation and illumination. The scaling problem is caused by the projection of the observed objects to the 2D visual sensing panels, making vision as an ill-posed problem. The rotation and translation problem arises from that the different position and orientation of objects may result into different appearances of the objects in the view. The illumination problem is caused by different light conditions in the environment and visual observations of objects may suffer from occlusions and shadows posed by the robot itself, particularly robot hands in grasping, and other objects in the scene.

In the past decades, vision sensors of different sensing principles have been proposed and commercialised, and many of them have been applied to the robotics research. The most widely used vision sensors are the RGB cameras. They are usually equipped with a standard CMOS sensor through which the color images of persons and objects are acquired. The acquisition of static photos is usually expressed in megapixels that means one million pixels, e.g., 2MP (1,920 x 1,080 = 2,073,600 pixels, also known as full HD resolution or 1080p), 12MP and 16MP. Compared to static images, videos captured by the RGB cameras can reveal the temporal information of the objects in the view, e.g., recognising human actions, tracking moving vehicles and localising a robot in a map.

To enable processing the visual events efficiently, event cameras, also known as neuromorphic cameras or or dynamic vision sensors, emerge in recent years. An event camera responds to local changes in brightness, instead of capturing images using a shutter as conventional cameras do. Pixels inside an event camera operates independently and asynchronously: each pixel reports changes in brightness as they occur and stays silent otherwise. Event cameras demonstrate better temporal resolution in order of millions fps (frames per second) compared to conventional cameras in order of hundreds fps.

Apart from 2D information extracted from images by the RGB cameras, 3D information can also be captured by vision sensors. One natural way to obtain the 3D information is to simulate the binocular vision of humans that derives information about how far away objects are based on solely relative positions of the object in the two eyes. A stereo camera simulates the human binocular vision by having two image sensors and therefore gives it the ability to perceive depth. There are also other ways to obtain the depth information based on different techniques, e.g., Time-of-Flight (ToF), structured light, and light fields. These depth sensors are usually used with the RGB cameras to form the RGB-D cameras so that both 2D appearances cues and depth can be obtained at the same time.

In the recent years, vision sensors have also been used in other sensing modalities. There is a rapid development of embedding vision sensors in optical tactile sensors. Such sensors usually consist of a visual camera at the base of the sensor and an elastomer layer on the top to interact with objects, and the visual

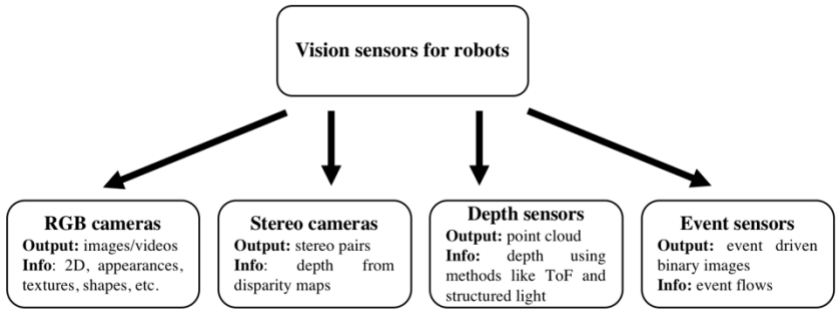

Figure 1.1 There are different types of vision sensors for robotics applications. RGB cameras have been one of the most widely used sensors in robotics, from robot grasping to visual SLAM. With the images or videos captured by the RGB cameras, rich information of the objects in the scene can be obtained, e.g., appearances, textures and shapes. Stereo cameras can be used to obtain the depth from the object to the camera from the obtained stereo pairs. Other depth sensors include ones based on ToF and structured light. Other types of vision sensors have also emerged and have been applied to robotics like event sensors that output event flows.

camera can capture the deformation of the elastomer in the interaction. The optical tactile sensors bridge the gap between vision and tactile sensing to create cross-modal perception. As visual cameras are used to capture the tactile interactions, the outputs of the optical tactile sensors are essentially camera images. This crossover has enabled techniques developed for computer vision, e.g., convolutional neural networks, to be applied to tactile sensing, connecting the look and feel of objects being interacted with. Recently there is also development that can transform between or match visual and tactile data from such sensors.

In this chapter, we introduce different vision sensors, i.e., RGB cameras, stereo cameras and depth sensors, that have been used in the robotics research, with an overview shown in Figure 1.1. We will then introduce how vision sensors can be used in other sensing modalities. Various aspects of vision sensors for robotic applications will be covered, including the hardware, integration, computation algorithms and applications to robotics.

1.2 RGB cameras for robotic perception

The projections of real 3D scenes onto 2D planes, generated when light (rays) real objects and filtered through a small cavity has always amused and served as a practical tool to humans, as pointed out by speculative theories about how prehistoric man produced cave paintings and the usage of camera obscura[3] in ancient and more modern civilisations. More recently, this working principle has been at the core of modern, firstly analogue and then digital cameras. This basic

[3] https://en.wikipedia.org/wiki/Camera_obscura

working principle enables cameras to capture a scene in an energy efficient manner and instantly.

While the real light phenomenon generates inverted projections, we can conceptually solve this by considering a virtual plane between the observed scene and the camera plane. Given the similarity between the two triangles, the image projected on the virtual plane is proportionally equivalent to the one projected on the real plane. By making the usual thin lens assumptions, the optical sensor can be modeled as a pinhole camera. The projective transformation that maps a point in the world space P into a point in a camera image P' can be defined using the general camera model [2] as:

$$P' = K[\boldsymbol{R}|\boldsymbol{t}]P$$

$$K = \begin{bmatrix} fk & 0 & c_x \\ 0 & fl & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

where $P' = [x'z, y'z, z]^T$ is an image pixel and $P = [x, y, z, 1]^T$ is a point in space, both represented in homogeneous coordinates here, $[\boldsymbol{R}|\boldsymbol{t}]$ is the camera's extrinsic matrix that encodes the rotation R and translation t of the camera, K is the camera intrinsic matrix (f is the focal length; k and l are the pixel-to-meters ratios; c_x and c_y are the offsets in the image frame). If the used camera produces square pixels, i.e., $k = l$, fk and fl can be replaced by α, for mathematical convenience. From the above equations, a point in the world space P can be mapped into a point in an image P' which is a "well posed" problem, i.e., has a uniquely determined solution. However, the mapping from P' to P usually does not have a uniquely determined solution, i.e., an "ill posed" problem. It results into the fact that a camera image may be resulted from different real-world settings. As a result, it is challenging for a robot to understand its ambient world from one single camera image.

1.3 Stereo Cameras

Given that cameras reduce 3D geometry into 2D this creates one problem: the loss of depth perception and/or size ambiguity. To mitigate this, two cameras and projections can be considered instead, to form a stereo camera. One of the commercially available stereo cameras is the ZED sensor[4].

By performing triangulation between the real point and the two corresponding projections, the depth of the object can be inferred. This construction is commonly referred as epipolar geometry. By comparing information about a scene from two corresponding points in left and right cameras that are projected from the same real-world point, 3D information can be extracted by examining the relative positions of objects in the two panels of the left and right cameras. The challenge in forming the epipolar geometry falls in matching the corresponding points in the two cameras, i.e., stereo matching. A large number of algorithms have been proposed for stereo correspondence using convolutional neural networks in recent years [3,4]. It has great potential to have stereo vision for robotics tasks as well, for example grasping.

1.4 Event cameras

[4] https://www.stereolabs.com/zed/

1.1.1 Event cameras - hardware

Compared with conventional cameras, the event cameras offer a number of advantages, including lower latency, less power, microsecond temporal resolution and larger dynamic range. Different from conventional cameras, the event camera employs a Dynamic Vision Sensor (DVS), which is able to capture the changes of brightness for each pixel asynchronously [5]. As a result, the event camera provides an asynchronous stream of brightness changes including the location, time information and polarity ("On" and "Off"), i.e., events. In DVS, each pixel memorises the statement of brightness as a reference when an event is triggered, and compares it with the current statement. If there exits an obvious variation that surpasses the threshold, an event is triggered and the reference is updated by the pixel.

On the other side, due to the use of "On" and "Off" polarity, it is difficult to construct a clear and detailed description of a scene. To address this problem, Asynchronous Time Based Image Sensor (ATIS) [6] and Dynamic and Active Pixel Vision Sensor (DAVIS) [7] have been proposed for event cameras. The ATIS includes the DVS pixels for the brightness change detection and another subpixel to measure the absolute values of brightness. As a result, ATIS can capture not only the motion but the background of static scene. The DAVIS consists of a DVS and a conventional Active Pixel Sensor (APS). This combination makes it be able to generate the colorful and detailed static background. However, the APS image usually suffers from motion blur and it is difficult to synchronise with DVS in high-speed motion scenes.

1.1.2 Event cameras - applications in robotics

In recent years, event cameras have been widely used in many robotic applications, such as object tracking [8,9], optical flow estimation [10,11], 3D reconstruction [12,13] and recognition tasks [14-16]. Compared to the conventional vision sensors, event cameras demonstrate nice features of low latency, less power and temporal resolution. These features make event sensors highly suitable for tasks that have strong requirements of efficiency in visual processing.

There are also advancements in the algorithmic development of the event based visual processing in the recent years. The Spiking Neural Networks (SNNs) has become a popular method for processing event signals. In the neurons of SNNs, the input signals, i.e., events, are received by the neurons and accumulated in the internal state, named as "membrane potential". When it exceeds a threshold, the neuron generates a spike for the neurons of the next layer and the internal state of the current neuron resets. Variants of SNNs have been developed such as Leaky Integrate and Fire models (LIF) and Spike Response Model (SRM) [17], inspired by the properties of human neurons. Thanks to this event-based property, the SNNs have been widely applied with the event camera in many applications, such as [18-20].

1.5 Depth cameras

Like its name suggests, RGB-D cameras are able to augment the RGB image with depth information, i.e., the distance from each point in the real scene to the

cameras. With the ability to measure object depth, RGB-D cameras have been widely used for object pose estimation, 3D reconstruction, and robotic grasping. In order to adapt to different application scenarios, many consumer-grade depth cameras have been developed in recent years.

According to the sensor types used in the cameras, RGB-D cameras can be divided into two categories, i.e., optical depth cameras and non-optical depth cameras. Optical depth cameras occupy a major part of the market thanks to its mature technology, low price and compact size. The most common techniques currently being employed for optical RGB-D cameras are based on structured lights, Time of Flight (ToF), and active Infrared (IR) Stereo methods.

Structured-lights based RGB-D cameras have a pair of a near-infrared laser transmitter and a receiver. It uses the transmitter and the receiver to project the light with certain structural features onto the object and collect the reflected light signals, respectively. Then it calculates the depth information based on the changes in the reflected light signals caused by different depth areas. In the early stages of RGB-D camera development, structured-lights RGB-D cameras attracted attention due to its mature technology, low cost, and low resource consumption. Some iconic examples of structured-lights based RGB-D cameras are Intel RealSense R200[5] and Microsoft Kinect V1[6]. However, structured-lights based RGB-D cameras are easily affected by the ambient light and long perception distance, which makes it not suitable for outdoor and large scenes.

Thanks to the increasing processing power, Microsoft successively launched two ToF based RGB-D cameras, Kinect V2 in 2014 and Azure Kinect[7] in 2018. Different from estimating depth with light signal changes in structure-lights based method, ToF based RGB-D cameras obtain the distance of the target by detecting the round-trip time of the light pulse. Through this way, it can work for long distance detection and reduce the interference of the ambient light. Nonetheless, the larger size of the ToF based RGB-D cameras limits their use on small mobile platforms like in many robotics applications.

In addition to the cameras mentioned above, RGB-D cameras based on the active IR stereo principle have also played an important role in the development of reliable depth sensors. Different from the naive block-matching methods that are widely used in stereo vision, the active IR stereo cameras use an infrared laser projector to generate texture for the stereo cameras, which significantly improves the accuracy. Moreover, the projector can be used as an artificial source of light for nighttime or dark situations. There are different series of RGB-D cameras based on active IR stereo technology such as Intel RealSense D415[8], D435[9], D435i[10], and D455[11].

1.6 Vision sensors for other modalities

[5] https://software.intel.com/content/www/us/en/develop/articles/realsense-r200-camera.html
[6] https://en.wikipedia.org/wiki/Kinect
[7] https://azure.microsoft.com/en-gb/services/kinect-dk/
[8] https://www.intelrealsense.com/depth-camera-d415/
[9] https://www.intelrealsense.com/depth-camera-d435/
[10] https://www.intelrealsense.com/depth-camera-d435i/
[11] https://www.intelrealsense.com/depth-camera-d455/

Cameras (and depth-cameras) can be used to assess large areas instantly, however, these suffer from occlusions and variances in the scene illumination, shadows and other sources of ambiguities. In contrast, tactile sensors offer a local assessment of the scene that is robust to such problems and, given the fact that precise sensing is more critical near contact, tactile sensors become a crucial sensing modality to consider, that is complementary to vision. Nonetheless, the fabrication of tactile skins is widely challenging, due to complicated electronics, cross-talk problems and consequently have traditionally produced low resolution of tactile signals [21-24]. On the other hand, cameras are these days ubiquitous, and consequently have become extremely cheap while being able to capture high-resolution images. As a consequence, a wide range of works have focused on exploiting such high-resolution cameras to produce optical tactile sensors.

Optical tactile sensors can be grouped in two main groups: marker-based and image-based, with the former being pioneered by the TacTip sensors [25] and the latter by the GelSight sensors [26]. As the name suggests, marker-based sensors exploit the tracking of markers printed on a soft domed membrane to perceive the membrane displacement and the resulted contact forces. By contrast, image-based sensors directly perceive the raw membrane with a variety of image recognition methods to recognise textures, localise contacts and reconstruct the membrane deformations, etc. Because of the different working mechanisms, marker-based sensors measure the surface on a lower resolution grid of points, whereas image-based sensors make use of the full resolution provided by the camera. Some GelSight sensors have also been produced with markers printed on the sensing membrane [27], enabling marker-based and image-based methods to be used with the same sensor. Both families of sensors have been produced with either flat sensing surfaces or domed/finger-shaped surfaces.

1.1.3 Marker-based sensors

The first marker-based sensor proposal can be found in [28], however more recently an important family of marker-based tactile sensors is the TacTip Family of sensors described in [29]. Since its initial domed shaped version [25], different morphologies have been proposed: including the TacTip-GR2 [30], a smaller fingertip design, TacTip-M2 [31], mimicking a large thumb for in-hand linear manipulation experiments, and TacCylinder to be used in capsule endoscopy applications. With its miniaturised and adapted design, [30,31] have been successfully used as fingers (or finger tips) in robotic grippers. Although each TacTip sensor introduces some manufacturing improvements or novel surface geometries, the same working principle is shared: white pins are imprinted onto a black membrane that can then be tracked using computer vision methods.

There are also other optical tactile sensors that track the movements of markers. In [32], an optical tactile sensor named FingerVision is proposed to make use of a transparent membrane, with the advantage of gaining proximity sensing. However, the usage of the transparent membrane makes the sensor lack the robustness to external illumination variance associated with touch sensing. In [33], semi-opaque grids of magenta and yellow makers, painted on the top and bottom surfaces of a transparent membrane are proposed, in which the mixture of the two colours is used to detect horizontal displacements of the elastomer. In [34], green florescent particles are randomly distributed within the soft elastomer with black

opaque coating so that a higher number of markers can be tracked and used to predict the interaction with the object, according to the authors. In [35] a sensor with the same membrane construction method, 4 Raspberry PI cameras and fisheye lenses has been proposed for optical tactile skins. A summary of influential Marker-based optical tactile sensors is shown in Table 1.1.

Table 1.1 A summary of influential Marker-based optical tactile sensors

	Sensor Structures	**Illumination and Tactile Membrane**
TacTip [25]	A domed (finger) shape, 40 × 40 × 85 mm^3, and tracks 127 pins; uses a Microsoft LifeCam HD webcam.	The membrane is black on the outside, with white pins and filled with transparent elastomer inside. Initially the membrane was cast from VytaFlex 60 silicone rubber, the pins painted by hand and the tip filled with the optically clear silicone gel (Techsil, RTV27905); however, currently the entire sensor can be 3d-printed using a multi-material printer (Stratasys Objet 260 Connex), with the rigid parts printed in Vero White material and the compliant skin in the rubber-like TangoBlack+.
TacTip-M2 [31]	A thumb-like or semi-cylindrical shape, 32 × 102 × 95 mm^3 and tracks 80 pins.	
TacTip-GR2 [30]	A cone shape with a flat sensing membrane, 40 × 40 × 44 mm^3, tracks 127 pins and uses an Adafruit SPY PI camera.	
TacCylinder [36]	A catadioptric mirror is used to track the 180 markers around the sensor cylindrical body.	
FingerVision [32]	It uses a ELP Co. USBFHD01M-L180 camera with a 180 degree fisheye lens. It has approximately 40 × 47 × 30 mm.	The membrane is transparent, made with Silicones Inc. XP-565, with 4 mm of thickness and markers spaced by 5 mm. No internal illumination is used, as its membrane is transparent
Subtractive Color Mixing [33]	N/A	Two layers of semi-opaque colored markers is proposed. Sorta-Clear 12 from Smooth-On, clear and with Ignite pigment, is used to make the inner and outer sides.
Green Markers [34]	The sensor has a flat sensing surface, measures 50 × 50 × 37 mm and is equipped with a ELP USBFHD06H RGB camera with a fisheye lens.	It is composed of three layers: stiff elastomer, soft elastomer with randomly distributed green florescent particles in it and black opaque coating. The stiff layer is

Multi-camera Skin [35]	It has a flat prismatic shape of 49 × 51 × 17.45 mm. Four Pi cameras are assembled in a 2 × 2 array and fish-eye lenses are used to enable its thin shape.	made of ELASTOSIL® RT 601 RTV-2 and is poured directly on top of the electronics, the soft layer is made of Ecoflex™ GEL (shore hardness 000-35) with the markers mixed in, and the final coat layer is made of ELAS- TOSIL® RT 601 RTV-2 (shore hardness 10A) black silicone. A custom board with an array of SMD white LEDs is mounted on the sensor base, around the camera.

1.1.4 Image based sensors

On the other side of the spectrum, the GelSight sensors, initially proposed in [26], exploit the entire resolution of the tactile images captured by the sensor camera, instead of just tracking makers. Due to the soft opaque tactile membrane, the captured images are robust to external light variations, and capture information of the touched surface's geometry structure, unlike most conventional tactile sensors that measure the touching force. Leveraging the high resolution of the captured tactile images, high accuracy geometry reconstructions are produced in [37-40]. In [37], this sensor was used as fingers of a robotic gripper to insert a USB cable into the correspondent port effectively. However, the sensor only measures a small flat area oriented towards the grasp closure.

Markers were also added to the membrane of the GelSight sensors, enabling applying the same set of methods that were explored in the TacTip sensors. There are some other sensor designs and adaptations for robotic fingers in [41-43]. In [41], matte aluminium powder was used for improved surface reconstruction, together with the LEDs being placed next to the elastomer, and the elastomer being slightly curved on the top/external side. In [42], the GelSlim is proposed, a design wherein a mirror is placed at a shallow and oblique angle for a slimmer design. The camera was placed on the side of the tactile membrane, such that it captures the tactile image reflected onto the mirror. A stretchy textured fabric was also placed on top of the tactile membrane to prevent damages to the elastomer and to improve tactile signal strength. Recently, an even more slim design has been proposed 2 mm [44], wherein a hexagonal prismatic shaping lens is used to ensure radially simetrically illumination. In [43], DIGIT is also proposed, an ease to manufacture and use sensor, with a USB “plug-and-play” port and an easily replaceable elastomer secured with a single screw mount.

In these previous works on camera based optical tactile sensors, multiple designs and two distinct working principles have been exploited. However, none of these sensors has the capability of sensing the entire surface of a robotic finger, i.e., both the sides and the tip of the finger. As a result, they are highly constrained in object manipulation tasks, due to the fact that the contacts can only be sensed when

the manipulated object is within the grasp closure [37,45]. To address this gap, we propose the fingertip-shaped sensor named GelTip that captures tactile images by a camera placed in the center of a finger-shaped tactile membrane. It has a large sensing area of approximately 75 cm^2 (*vs.*) 4 cm^2 of the GelSight sensor) and a high resolution of 2.1 megapixels over both the sides and the tip of the finger, with a small diameter of 3 cm (vs. 4 cm of the TacTip sensor). More details of the main differences between the GelSight sensors, TacTip sensors and our GelTip sensor are given in Table 1.2.

With its compact design, the GelTip [46,47] and other GelSight [37, 42-44] sensors are candidate sensors to be mounted on robotic grippers, however custom grippers and sensors built using the GelSight working principle have also been proposed [48,49]. Simulation models of such sensors have also been proposed [50, 51].

Two recent works [52,53] also address the issue of the flat surface of previous GelSight sensors. However, their designs have large differences to ours. In [52], the proposed design has a tactile membrane with a surface geometry close to a quarter of a sphere. Therefore, a great portion of contacts happening on the regions outside the grasp closure is undetectable. In [53], this issue is mitigated using five endoscope micro cameras looking at different regions of the finger. However, this results in a significant increase of cost for the sensor, according to the authors, approximately US$3200 vs. only around US$100 for ours).

Table 1.2 A summary of influential flat and finger-shaped GelSight sensors

	Sensor Structures	**Illumination**	**Tactile Membrane**
GelSight [37]	It has a cubic design with a flat square surface. A Logitech C310 (1280 × 720) camera is placed at its base pointing at the top membrane.	Four LEDs (RGB and white) are placed at the base. The emitted light is guided by the transparent hard surfaces on the sides, so that it enters the membrane tangentially.	A soft elastomer layer is placed on top of a rigid, flat and transparent acrylic sheet. It is painted using semi-specular aluminum flake powder.
GelSight [41]	It has a close-to hexagonal prism shape. The used webcam is also the Logitech C310.	Three sets of RGB LEDs are positioned (close to) tangent to the elastomer, with a 120° angle from each other.	A matte aluminium powder is proposed for improved surface reconstruction. Its elastomer has a flat bottom and a curved top.
GelSlim [42]	A mirror placed at a shallow oblique angle and a Raspberry Pi Spy (640 × 480) camera is used to capture the tactile image reflected by the mirror.	A single set of white LEDs is used. These are pointed at the mirror, so that the light is reflected directly onto the tactile membrane.	A stretchy and textured fabric on the tactile membrane prevents damages to the elastomer and results in improved tactile signal strength.

GelSlim v3 [44]	It is shaped similar to [37, 41] however slimmer 20 mm of thickness, and a round sensing surface.	A custom hexagonal prism is constructed to ensure radially symmetric illumination.	An elastomer with Lambertian reflectance is used, as proposed in [41].
DIGIT [43]	A prismatic design, with curved sides. An OmniVision OVM7692 (640 × 480) camera is embedded in the custom circuit board.	Three RGB LEDs are soldered directly into the circuit board, illuminating directly the tactile membrane.	The elastomer can be quickly replaced using a single screw mount.
Round Fingertip [50]	It has a round membrane, close to a quarter of sphere. A single 160° FoV Raspberry Pi (640 × 480) is installed on its base.	Two rings of LEDs are placed on the base of the sensor, with the light being guided through the elastomer.	Both rigid and soft parts of the membrane are cast, using SLA 3D printed molds.
OmniTact [51]	It has a domed shape. Five endoscope cameras (400 × 400) are installed on a core mount, and placed orthogonally to each other: pointing at the tip and sides.	RGB LEDs are soldered both onto the top and sides of the sensor.	The elastomer gel is directly poured onto the core mount (and cameras) without any rigid surface or empty space in between.
GelTip [46,47]	It has a domed (finger) shape, similar to a human finger. A Microsoft Lifecam Studio webcam (1920× 1080) is used.	Three sets of LEDs, with a 120° angle from each other, are placed at the sensor base, and the light is guided through the elastomer	An acrylic test tube is used as the rigid part of the membrane. The deformable elastomer is cast using a three-part SLA/FFF 3D printed mold.

1.7 Conclusions

In this chapter, we introduced different aspects of the vision sensors for robotics applications, from their working principles to their applications to robotics, particularly on robot perception and their use in sensors of other modalities. As one of the most widely used sensors for robots, they have advantages of low cost and high resolution. Vision sensors have also been used in other sensing modalities and we have introduced the state-of-the-art research in optical tactile sensors using visual sensors. By having the vision sensors, robots can sense and estimate the properties of the objects that they interact with, e.g., object shapes, appearances, textures and mechanical parameters. It can be forecast that vision sensors will be one of the most widely used sensors in the research of robotics, and new types of vision sensors, like event sensors and more robust RGB-D sensors, will emerge in the future research and development.

1.8 Acknowledgements

This work was supported by the EPSRC project "ViTac: Visual-Tactile Synergy for Handling Flexible Materials" (EP/T033517/1).

1.9 References

1. B. R. Sheth, J. Sharma, S. C. Rao, M. Sur, Orientation maps of subjective contours in visual cortex, Science 274 (5295) (1996) 2110–2115.
2. R. Szeliski, Computer vision algorithms and applications, Springer Science & Business Media, 2010.
3. J. Zbontar, Y. LeCun, et al., Stereo matching by training a convolutional neural network to compare image patches., J. Mach. Learn. Res. 17 (1) (2016) 2287–2318.
4. J. Zbontar, Y. LeCun, Computing the stereo matching cost with a convolutional neural network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1592–1599.
5. P. Lichtsteiner, C. Posch, T. Delbruck, A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor, IEEE Journal of Solid-State Circuits 43 (2) (2008) 566–576. doi:10.1109/JSSC.2007.914337.
6. C. Posch, D. Matolin, R. Wohlgenannt, A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds, IEEE Journal of Solid-State Circuits 46 (1) (2010) 259–275.
7. C. Brandli, R. Berner, M. Yang, S.-C. Liu, T. Delbruck, A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor, IEEE Journal of Solid-State Circuits 49 (10) (2014) 2333–2341.
8. B. Ramesh, S. Zhang, Z. W. Lee, Z. Gao, G. Orchard, C. Xiang, Long-term object tracking with a moving event camera., in: Bmvc, 2018, p. 241.

9. T. Delbruck, M. Lang, Robotic goalie with 3 ms reaction time at 4% cpu load using event-based dynamic vision sensor, Frontiers in neuroscience 7 (2013) 223.
10. M. Liu, T. Delbruck, Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors (2018).
11. A. Z. Zhu, L. Yuan, K. Chaney, K. Daniilidis, Ev-flownet: Self-supervised optical flow estimation for event-based cameras, arXiv preprint arXiv:1802.06898 (2018).
12. H. Rebecq, G. Gallego, E. Mueggler, D. Scaramuzza, Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time, International Journal of Computer Vision 126 (12) (2018) 1394–1414.
13. H. Kim, S. Leutenegger, A. J. Davison, Real-time 3d reconstruction and 6-dof tracking with an event camera, in: European Conference on Computer Vision, Springer, 2016, pp. 349–364.
14. R. Ghosh, A. Mishra, G. Orchard, N. V. Thakor, Real-time object recognition and orientation estimation using an event-based camera and cnn, in: 2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings, IEEE, 2014, pp. 544–547.
15. A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, et al., A low power, fully event-based gesture recognition system, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7243–7252.
16. Q. Wang, Y. Zhang, J. Yuan, Y. Lu, Space-time event clouds for gesture recognition: From rgb cameras to event cameras, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1826–1835.
17. W. Gerstner, W. M. Kistler, Spiking neuron models: Single neurons, populations, plasticity, Cambridge university press, 2002.
18. T. Taunyazov, W. Sng, H. H. See, B. Lim, J. Kuan, A. F. Ansari, B. C. Tee, H. Soh, Event-driven visual-tactile sensing and learning for robots, arXiv preprint arXiv:2009.07083 (2020).
19. R. Massa, A. Marchisio, M. Martina, M. Shafique, An efficient spiking neural network for recognizing gestures with a dvs camera on the loihi neuromorphic processor, arXiv preprint arXiv:2006.09985 (2020).
20. M. Gehrig, S. B. Shrestha, D. Mouritzen, D. Scaramuzza, Event-based angular velocity regression with spiking networks, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, pp. 4195–4202.
21. S. Luo, W. Mou, K. Althoefer, H. Liu, Localizing the object contact through matching tactile features with visual map. in: 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 3903-3908.
22. S. Luo, W. Mou, K. Althoefer, H. Liu, Novel Tactile-SIFT descriptor for object shape recognition, IEEE Sensors Journal 15 (9) (2015) 5001–5009.
23. S. Luo, W. Mou, K. Althoefer, H. Liu, Iterative closest labeled point for tactile object shape recognition, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 3137–3142.
24. S. Luo, W. Mou, K. Althoefer, H. Liu, iCLAP: Shape recognition by combining

proprioception and touch sensing, Autonomous Robots 43 (4) (2019) 993–1004.
25. C. Chorley, C. Melhuish, T. Pipe, J. Rossiter, Development of a tactile sensor based on biologically inspired edge encoding, International Conference on Advanced Robotics (ICAR) (2009).
26. M. K. Johnson, E. H. Adelson, Retrographic sensing for the measurement of surface texture and shape Retrographic sensing for the measurement of surface texture and shape, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009. doi:10.1109/CVPR.2009.5206534.
27. S. Dong, W. Yuan, E. H. Adelson, Improved GelSight Tactile Sensor for Measuring Geometry and Slip, CoRR abs/1708.0 (2017).
28. K. Vlack, K. Kamiyama, T. Mizota, H. Kajimoto, N. Kawakami, S. Tachi, Gelforce: A traction field tactile sensor for rich human-computer interaction, in: IEEE Conference on Robotics and Automation, 2004. TExCRA Technical Exhibition Based., 2004, pp. 11–12. doi:10.1109/TEXCRA.2004.1424969.
29. B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, N. F. Lepora, The TacTip Family: Soft Optical Tactile Sensors with 3D-Printed Biomimetic Morphologies, Soft Robotics 5 (2) (2018) 216–227.
30. B. Ward-Cherrier, N. Rojas, N. F. Lepora, Model-free precise in-hand manipulation with a 3d-printed tactile gripper, IEEE Robotics and Automation Letters 2 (4) (2017) 2056–2063. doi:10.1109/LRA.2017.2719761.
31. B. Ward-Cherrier, L. Cramphorn, N. F. Lepora, Tactile manipulation with a tac-thumb integrated on the open-hand m2 gripper, IEEE Robotics and Automation Letters 1 (1) (2016) 169–175.
32. A. Yamaguchi, C. G. Atkeson, Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables, in: IEEE-RAS International Conference on Humanoid Robots, 2016, pp. 1045–1051. doi:10.1109/HUMANOIDS.2016.7803400.
33. X. Lin, M. Wiertlewski, Sensing the Frictional State of a Robotic Skin via Subtractive Color Mixing, IEEE Robotics and Automation Letters 4 (3) (2019) 2386–2392. doi:10.1109/LRA.2019.2893434.
34. C. Sferrazza, R. D'Andrea, Design, motivation and evaluation of a full-resolution optical tactile sensor, Sensors 19 (4) (2019).
35. C. Trueeb, C. Sferrazza, R. D'Andrea, Towards vision-based robotic skins: a data-driven, multi-camera tactile sensor, in: 2020 3rd IEEE International Conference on Soft Robotics (RoboSoft), 2020, pp. 333–338.
36. B. Winstone, C. Melhuish, T. Pipe, M. Callaway, S. Dogramadzi, Toward bio-inspired tactile sensing capsule endoscopy for detection of submucosal tumors, IEEE Sensors Journal 17 (3) (2017) 848–857.
37. R. Li, R. Platt Jr, W. Yuan, A. Pas, N. Roscup, M. A. Srinivasan, E. H. Adelson, Localization and Manipulation of Small Parts Using GelSight Tactile Sensing, IEEE International Conference on Intelligent Robots and Systems (2014).
38. S. Luo, W. Yuan, E. Adelson, A. G. Cohn, R. Fuentes, ViTac: Feature sharing between vision and tactile sensing for cloth texture recognition, in: IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 2722–2727.

39. J.-T. Lee, D. Bollegala, S. Luo, “touching to see” and “seeing to feel”: Robotic cross-modal sensory data generation for visual-tactile perception, in: IEEE International Conference on Robotics and Automation (ICRA), 2019, pp. 4276–4282.
40. G. Cao, Y. Zhou, D. Bollegala, S. Luo, Spatio-temporal attention model for tactile texture recognition, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020.
41. W. Yuan, S. Dong, E. H. Adelson, GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force., Sensors (Basel, Switzerland) 17 (12) (11 2017).
42. E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, A. Rodriguez, Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 1927–1934.
43. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation 5.
44. I. Taylor, S. Dong, A. Rodriguez, Gelslim3. 0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger, arXiv preprint arXiv:2103.12269 (2021).
45. S. Dong, D. Ma, E. Donlon, A. Rodriguez, Maintaining grasps within slipping bounds by monitoring incipient slip, in: IEEE International Conference on Robotics and Automation (ICRA), 2019, pp. 3818–3824.
46. D. F. Gomes, Z. Lin, S. Luo, GelTip: A finger-shaped optical tactile sensor for robotic manipulation, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020.
47. D. F. Gomes, Z. Lin, S. Luo, Blocks world of touch: Exploiting the advantages of all-around finger sensing in robot grasping, Frontiers in Robotics and AI 7 (2020).
48. G. Cao, J. Jiang, C. Lu, D.F. Gomes, S. Luo. TouchRoller, A rolling optical tactile sensor for rapid assessment of large surfaces. arXiv preprint arXiv:2103.00595 (2021).
49. Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, E. Adelson, Cable manipulation with a tactile-reactive gripper, arXiv preprint arXiv:1910.02860 (2019).
50. D.F. Gomes, A. Wilson, S. Luo, Gelsight simulation for sim2real learning. In ICRA ViTac Workshop, 2019.
51. D.F. Gomes, P. Paoletti, S. Luo, Generation of GelSight tactile images for sim2real learning. IEEE Robotics and Automation Letters, 6(2) (2021) pp.4177-4184.
52. B. Romero, F. Veiga, E. Adelson, Soft, Round, High Resolution Tactile Fingertip Sensors for Dexterous Robotic Manipulation, in: IEEE International Conference on Robotics and Automation, 2020.
53. A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, S. Levine, Omnitact: A multi-directional high-resolution touch sensor, in: IEEE International Conference on Robotics and Automation, 2020.