Tactile and Vision Perception for Intelligent Humanoids

Shuo Gao,* Yanning Dai, and Arokia Nathan*

Touch and vision perception are two important functions humans use to interact with the real world. To mimic human-like abilities, tactile- and visual-sensing-based intelligent humanoids have emerged and are going through a fast phase of development. Studies have demonstrated that the combination of tactile and visual information not only enables humanoids to better learn the environment, but also allows them to have pseudocognitive ability. Being a new and rapidly developing field of research, a significant growth in articles reporting different aspects of sensing and related machine learning is being witnessed. To help readers comprehensively understand the fundamentals and insights, and the current state of the art, this review is compiled to explain the working mechanisms of tactile and visual sensing, introduce the application of intelligent humanoids in diverse scenarios, discuss current challenges, and predict future trends.

1. Introduction

Over the last few decades, robotics development has evolved to mimic human appearance and behavior so as to enable smooth, efficient human–robotic collaboration (HRC) in complex and dangerous scenarios.[1,2] In this context, the concept of the humanoid, called Herbert Televox, was proposed in 1927.[3] Since then, various humanoids have traditionally been proposed, and used for conducting simple industrial tasks, such as assembling elements and distributing items.[4] Although considerable human power has been released, the aim of efficient HRC has yet to be achieved.

What makes it so hard? The present-day researches dominantly focus on replicating one or very few human actions, although failing to build human-like robotic intelligence. This intelligence mainly lies in four aspects: perception, cognition, motion, and interaction. Specifically, the humanoid robot should “feel” a vastly rich set of environmental information, synthesize and interpret it into cognitive clues, and then generate reasonable behaviors; these procedures are continually occurring for adjusting the interactive methods to the changing environment. Therefore, the challenges par excellence for achieving intelligence can be concluded as the difficulty in developing multidimensional and highly dependable sensing systems and establishing the relationship between origin sensory information and cognition/control signals.

In recent years, considerable advances in engineering and computer science have enabled rapid development of smart sensors and machine learning algorithms, giving rise to the emergence of intelligent humanoids, which were previously only depicted in science fiction. Intelligent humanoids have multiple sensors capable of detecting multidimensional environmental information, in which the data acquired are processed by comprehensive algorithms. Smart sensors are considered human sensory organs, and comprehensive algorithms function as the human brain. As such, intelligent humanoids replicate the physical and cognitive abilities of humans.[5]

Among smart sensors, tactile-based transduction and optical-based transduction have received significant attention compared to their counterparts. This is because touch and vision are the two very important functions used by humans to learn and interact with the real world.[6] Vision can quickly and directly discriminate an object’s location, shape, and color. Therefore, when humanoids are equipped with cameras, environment modeling and item identification is easily achieved.[7] In terms of tactile sensing, humanoids can detect contact-based information, such as roughness, texture, and weight, which can be further used for diverse purposes, for example, adjusting grasp position.[8]

Tactile sensing and vision sensing have different characteristics, and can be used for detecting information in distinct dimensions. Both can individually achieve tasks such as object recognition,[9,10] human behavior analysis,[11,12] and 3D information reconstruction.[13,14] However, as the knowledge acquired by using only the one or the other is still limited, they do not perform well when completing complex tasks. For example, it is challenging to identify large objects solely using tactile sensing.[15] Therefore, studies have focused on investigating use of
both tactile and vision information in a multisensory fusion method because their combination can retain their unique advantages and compensate for drawbacks.\textsuperscript{16,17} Experimental results have demonstrated that not only can the performance of conventional tasks for humanoids be improved,\textsuperscript{18–20} but also traditionally deemed impossible tasks can be completed.\textsuperscript{21–23} For example, estimation of the mass center of irregular-shaped items with uneven density is a difficult task for robotics. In fact, this is even true for humans. However, with the combination of visual and tactile information,\textsuperscript{21} the mass center position can be calculated with a high accuracy of 98.1\%.\textsuperscript{21} More application scenarios are shown in Figure 1.

As the integration of vision and tactile functions can endow a conventional humanoid with human abilities, an increasing number of researchers around the world are devoting themselves to the field, and their fruitful and promising outputs are steering new academic and industrial researchers to work toward this direction. However, this field covers a wide range of technologies, and currently a systematic review does not exist in the published literature, making it difficult for beginners to enter the field. Motivated by this, we provide in this article a comprehensive review on tactile- and vision-enabled intelligent humanoids, and notably, the algorithms that are essential to interpret sensor-acquired information. We do not specifically introduce each algorithm but instead show how algorithms can in general be used to compute sensory data.

This article is structured as follows. Sections 2 and 3 respectively introduce the underlying mechanisms and application scenarios of tactile and visual sensing in humanoids. Section 4 presents tactile and visual information fusion strategies, and describes advanced and novel functions provided by combining these two elements. Challenges and future development trends are addressed in Section 5.

2. Tactile-Enabled Humanoids

The principal idea of developing tactile sensors for humanoids is to mimic the touch function of humans. According to the human skin properties, as well as the results of several prior surveys to the researchers and manufacturers in the area of industrial robots, the basic robotic skin requirements for object recognition and manipulation applications can be summarized as the following points:\textsuperscript{24–26}: 1) force sensitivity of 0.05–0.1 N in the majority of tasks, and 0.005 N for some delicate manipulation tasks; 2) working range wider than 1000:1; 3) response time of 1–10 ms for each sensing element; 4) spatial resolution of about 1–2 mm; 5) good repeatability, low hysteresis, and monotonic output response (linearity is not necessary); 6) capabilities for dynamic and static force measurement; and 7) shear force or slip detection required in object handling, typically for limp materials.

To achieve these goals, diverse tactile sensing techniques based on distinct mechanisms have been reported in the literature. In this section, we will explain tactile sensors’ working principles, compare their pros and cons, and discuss corresponding applications. The flow diagram of this section is displayed in Figure 2.

2.1. Sensing Mechanisms

2.1.1. Capacitance

Capacitance-measurement-based tactile sensing normally focuses on detecting the contact force between humanoids and objects.\textsuperscript{27,28} To achieve this, elastic materials are normally inserted between two electrode layers. When a force event occurs between the humanoid and the object, the elastic layer is distorted, producing a change in capacitance.\textsuperscript{28,29}

Capacitive-based tactile sensors can detect normal force, shear force, and three-axis force. For detecting normal force, the elastic layer is compressed, shortening the distance between the two electrode layers. Parallel capacitors are the most widely used architecture for normal force detection.\textsuperscript{28} Due to the small deformation of the in-between elastomer, however, capacitive-based force-sensing techniques suffer from low-sensitivity and limited-measurement-range issues. To address this, Bao and co-workers first used microstructured patterns to the dielectric layers.\textsuperscript{29} With square pyramid arrays cast on a polydimethylsiloxane (PDMS) dielectric layer, the sensor’s sensitivity reached a 12 times enhanced sensitivity of 0.58 kPa\textsuperscript{1}. In addition, the microstructure surface provided voids for the sensor’s elastic deformation, thus minimizing the viscoelastic behaviors and reducing the response time. In recent studies, structuring the electrode layers rather than the dielectrics has also proved to be effective. For example, in Pang et al.,\textsuperscript{30} an Au/PEN electrode layer was fabricated with micro-pyramids and microhairs to improve sensing performance. The sensor offered a 12 times enhanced sensitivity of 0.58 kPa\textsuperscript{1}. Yang et al.\textsuperscript{31} used microcylinder arrays to graphene
electrodes and demonstrated a high sensitivity of 3.19 kPa$^{-1}$. In addition, the enhancement of the dielectric layer’s permittivity can also improve capacitive sensors’ sensitivity. By doping ionic additives to a polymer poly(vinylidene fluoride-co-hexafluoropropene) (PVDF-HFP) dielectric, Chen et al.$^{[32]}$ successfully improved the overall permittivity from 6.0 to 37.5, and obtained a 0.73 kPa$^{-1}$ sensitivity.

In terms of the shear force and three-axis force detection, a centrally symmetric capacitor array (usually a $2 \times 2$ matrix) can be used.$^{[27,33,34]}$ In which the normal force causes the same capacitance changes in the capacitor array, whereas the shear force causes different changes. In Liang et al.$^{[27]}$ a three-axis force sensor was proposed for robotic skin applications. Each sensor unit comprises a top PDMS bump that enhances the sensor sensitivity, a microstructured PDMS dielectric layer, and four pairs of square-shaped polyethylene terephthalate (PET) electrodes. When a force is applied on the top bump, the normal force component is calculated by the average capacitance change in the four sensor elements, and the shear component is deduced by the capacitance difference between adjacent elements. The device’s responsivities were tested to be 58.3%, 57.4%, and 67.2% per newton in the $x$, $y$, and $z$-axes within 0.5 N, respectively. In Boutry et al.$^{[13]}$ a triaxial smart skin with an interconnected structure was proposed, where the upper electrode was patterned with tiny pyramid arrays and the lower electrode was shaped with a larger hemispherical convex. In this architecture, an external force could deform at least a $5 \times 5$ pixel array. The changed capacitance data could be interpreted as the force amplitudes and orientations. The sensor showed a normal force sensitivity of 0.19 kPa$^{-1}$ and shear sensitivity of 3.0 kPa$^{-1}$.

Figure 2. Summary of commonly used tactile sensing mechanisms in humanoids, and their applications in different fields. Reproduced with permission.
In recent works, capacitive-based pressure-sensing transistors have been proposed for high-density integration and low-noise applications. For instance, in Schwartz et al., a field-effect transistor was reported for normal force sensing, which was constituted by a top indium tin oxide (ITO) gate electrode, a microstructured PDMS compressible dielectric film, a semiconducting polymer, and bottom interdigital source–drain electrodes. An applied force would change the capacity of the top capacitor, thus changing the transistor’s saturation drain current. Its maximum sensitivity achieved 8.4 kPa⁻¹ in the pressure range of <8 kPa; the sensitivity is 15-fold higher compared with conventional structures.

Capacitive tactile sensors provide the advantages of low power consumption, rapid response, and simple structure, but they are sensitive to EM noise. Elastic materials normally suffer from the effects of ageing, indicating that the detection accuracy is lowered over long-term use.

2.1.2. Piezoresistivity

The resistance values of piezoresistive materials are altered when force is applied. Widely used piezoresistive materials include conductive fabrics,36 conductive rubber,37 conductive foam,38 and ionic liquids.39,40 In Kovia et al., a fingertip-shaped sensor array was developed with a conductive elastomer foam and a copper electrode layer patterned by a laser-direct-structuring process. Its normal force-sensing range was measured to be 0.03–10 N. Noda et al. demonstrated a flexible robotic skin design with 1-ethyl-3 methylimidazolium ethylsulfate (EMIES) ionic liquid and a Parylen-C polymer film embedded with micro-channels, achieving a responsivity of 1.25% per newton within normal force of 5 N. In recent years, to enhance the piezoresistive effect of such sensors, nanocomposites and conductive nanomaterials have been widely demonstrated. For example, in Sun et al., a carbon nanotube (CNT)/PDMS nanocomposite was used as the sensing material, exhibiting a high sensitivity of 12.1 kPa⁻¹ in the regime within 0.6 kPa. In Chen et al., silver nanowires (AgNWs) and PEDOT:PSS are embedded in tissue paper materials to improve the materials’ piezoresistivity. The device’s sensitivity reached as high as 1089.7 kPa⁻¹ in the range of 0–1 kPa. Another enhancement solution is patterning micro-structures at the sensing layer. For example, Chen et al. patterned a piezoresistive layer with a ZnO nanorod array, and obtained high sensitivity at 88 kPa⁻¹ in a pressure regime of <10 kPa. Similar to the capacitive-based tactile sensors, piezoresistive architectures are also capable of detecting force in different orientations. For instance, a three-axis fingertip force sensor was proposed with an upper electrode layer composed of four symmetric sectors, a composite layer of an elastomer and carbon particles, and a lower electrode layer. The sensing ranges of normal and shear force were found to be 0.05–20 and 0.05–10 N, respectively.

In the last decade, piezoresistive sensors’ ability to directly convert applied pressure to resistance change has inspired scientists to develop various functional-material-based organic semiconductor devices. Sekitani et al. integrated a flexible floating-gate transistor array with piezoresistive sensors. The sensors acted as gate-source voltage adjusting resistances, which were able to shift the transistor’s threshold voltage. In addition, the applied force could be stored for at least 12 h after a power loss, potentially providing tactile memory functions for robot intelligent manipulations. In Chou et al., a chameleon-inspired color-changing smart skin was proposed, which comprised organic electrochromic devices (ECDs) and CNT-based piezoresistive force sensors. The applied forces (within 200 kPa) could be directly expressed by the device’s color changes. A piezoresistive transistor-based flexible circuit was proposed to directly convert pressure information within 0–100 kPa to frequency signals of 0–200 Hz. This conversion mimicked the characteristics of human skin mechanoreceptors and provided humanoid robots with realistic and quick-response neural reflex functions.

Piezoresistive tactile sensors provide the advantages of a simple structure and ease of large-area manufacture, but suffer from high hysteresis and temperature drift.

2.1.3. Piezoelectricity

Piezoelectric materials convert mechanical displacement into electric signal due to their intrinsic noncentrosymmetry architectures. Broadly used piezoelectric materials can be divided into ceramic, thin film, and fibric categories. Among them, ceramics have the highest piezoelectric coefficients, but are brittle during use. Hence, they are normally installed at the humanoid’s fingertip. For example, in Acer et al., a fingertip-type tactile sensor array was developed based on an electrode/PZT ceramic/electrode sandwiched structure to detect dynamic force and touch position. It achieved a 5 mm spatial resolution a 0.821 V/N responsivity under 0–1 N impulse forces. Piezoelectric thin films have good flexibility; therefore, they are suitable for covering a humanoid’s body area. For instance, in Goger et al., a PVDF film was used as the sensitive layer to sense the dynamic force induced by slipping in object manipulation applications. Using the machine learning techniques, the forces are connected to the slipping state, achieving an accuracy of 99.82% over 2600 experiments. Piezoelectric fibers are easily broken, so their use in humanoids has not yet been reported.

The piezoelectric responsivity can be further improved by importing microstructures or incorporating triboelectric effects. In Choi et al., a PZT piezoelectric layer was bonded with a microstructured PDMS enhancing film, exhibiting a wide dynamic stress range of 0.23–10 kPa. Zhao et al. integrated the piezoelectric effect and triboelectric effect by adding a negative triboelectric layer (PDMS) to a nanofiber-based piezoelectric sensor. Its force-sensing responsivity achieved 1.44 V N⁻¹ in the range of 0.15–20 N. Moreover, many studies report that the integration or cascade of piezoelectric sensors and transistors can provide lower electromagnetic noise, thus enhancing the sensor’s measurement capability. For example, an ultra-thin force-sensing robotic skin was proposed comprising a silicon flexible substrate, a single metal oxide semiconductor field-effect transistor (MOSFET), and a PZT force-sensing array connected to the transistor’s gate electrode in parallel. A high-level force sensitivity of 0.005 Pa was experimentally demonstrated. A ZnO piezoelectric layer was included in the gate stacks of thin-film transistors (TFTs) to form a force-sensitive piezo-TFT. In this way, the sensor and the analog circuitry were
combined in the same entity, allowing reduced interconnections and a high force responsivity of 0.207 % per micronewton.

Piezoelectric tactile sensors have good dynamic force-sensing performance, high sensitivity, and simple structure. However, due to their inability to sense static force, the integration of other sensors is required.

2.1.4. Others

In addition to the aforementioned mechanisms, several less reported tactile sensing techniques are introduced here.

The impedance-based tactile technique, i.e., electrical impedance tomography (EIT), installs several electrodes around a certain area. Through sending and receiving electric currents between electrodes, the impedance distribution information within the area is reconstructed. Tawil et al. presented a large-area flexible tactile sensor on a robotic forearm for force and touch position sensing. The sensor is composed of two conductive fabric layers of different materials and 19 boundary electrodes. The impedance distribution was reconstructed with the generalized Tikhonov regularization algorithm, and used for touch modality classification (e.g., tap, pat, push, and slap), finally reaching a 70.7% accuracy among nine modalities. EIT-based tactile sensing is applicable in large-area tactile sensing scenarios. However, EIT can only yield qualitative results, indicating a low detection accuracy.

Optical-based tactile sensors are able to detect normal force based on the light reflection principle. For example, a tactile sensor array was proposed on robotic arms to perform normal force perception. Each sensing element consisted of a deformable elastic shell, an inside light emitting diode (LED) light transmitter, and an inside phototransistor light receiver. When an external force deforms the shell, the light reflection condition inside is changed, thus changing the light density received by phototransistors. The sensor achieved a spatial resolution of 2 mm and sensitivity of 0.05 N in the range of 0–8 N. The optical sensors are immune to EMI; however, they suffer from complex structure and low durability.

Electromagnetic-induction-based techniques can measure three-axis forces by sensing the deformation-induced magnetic flux density variation. In Wattanasarn et al., a flexible 3D tactile sensor was proposed, which covered curved skin surfaces on robots. The sensor was structured as four parallel layers: a top bump, a detection layer, a spacer, and an excitation layer; both the detection and the excitation layers were embedded with four coils to sense the magnetic field changes. After a calibration process, the sensor showed root-mean-square errors (RMSEs) of 0.32 N, 0.47 N, and 14° for normal force, shear force, and force direction sensing, respectively. The electromagnetic induction sensors have high force sensitivity but low durability and low resistance to EMI.

Apart from the force-sensing techniques, temperature- and humidity-based tactile sensors have been reported. For example, a pressure–temperature dual-mode smart skin was proposed based on piezoresistive and thermoelectric-based techniques, where temperature differences on the sensor’s upper and lower surfaces would generate a voltage signal. The sensor exhibited a temperature gradient sensing range of 0–40°C and resolution of 0.1°C. This sensing ability was helpful to detect objects’ thermal conductivities in object recognition tasks. The humidity sensing is not a widely integrated function for tactile modules. It has been reported in the development of electric skins that monitor the environment information, but it is less used in object recognition or manipulation. There are three approaches to measure relative humidity (RH): capacitive-based, resistive-based, and triboelectric-based techniques. A capacitive-based humidity sensor structured as polyimide/copper electrodes/PVDF was proposed for relative humidity sensing on robot hands. Its working principle is when the environmental humidity changes, the polyimide layer absorbs or desorbs water vapor, alerting the polyimide’s permittivity, which is detected by the capacitor formed by planar electrodes. The device reached a humidity responsivity at 0.22%/RH% within a humidity range of 10–90% RH. For resistive-based devices, nanomaterials with large amounts of exposed atoms and high porosity are used for the hydration or dehydration of water molecules. This process would affect the spacing of material particles and change the overall conductivity. In Jeong et al. metal carbide/carbonitride nanosheets and AgNWs served as the porous humidity-sensing material, exhibiting a responsivity of 0.7%/RH% under the 5–80% RH environment. A resistive sensor composed of adhesive paper material and polyimide tapes was proposed, showing a high humidity-sensing response of more than three orders in the range of 41.1–91.5% RH. Triboelectric-based methods have been recently proposed to conduct humidity measurements with low power consumption. Similar to the resistive types, the triboelectric sensor’s internal resistance changes with the environment humidity, and transforms the output voltage levels according to a voltage-divider principle. For example, in Bahuguna et al. a triboelectric humidity sensor was proposed based on a copper friction layer, a polytetrafluoroethylene (PTFE) friction layer, and two tin disulfide nanoflower/reduced graphene oxide (SnS$_2$/RGO) humidity-sensing layers. When the two friction layers get to contact and separate, a steady electric output will generate, the value of which depends on the humidity layer’s resistance. The sensor offered a 65 times response in a wide humidity range of 0–97% RH experimentally.

In addition to properties of multidimension, high sensitivity, and large area, another strongly desired attribute for robotic tactile sensors is stretchability. It allows robotic skins to perfectly cover the machine body, especially at the moving parts like joints, on which the surface skin should allow stretches of more than 55%. There are three common strategies to achieve stretchability in electronic skins: 1) buckling the substrate surface, 2) connecting rigid islands with stretchable electric wires, and 3) using stretchable materials for all components, including electrodes and functional layers. For the first, electronic devices are attached to a prestrained elastomer substrate. When the substrate is released, buckled surface patterns are induced. This process can transform various flexible devices into stretchable ones; however, uneven issues prevent the sensor from intimate contact with robot surfaces. The second method, which uses curved or stretchable electric wires to connect stiff sensor arrays, has been extensively used in practical applications. For example, in Hua et al. a matrix network of pressure, temperature, humidity, and touch sensors is interconnected by wires to achieve large
expansion (300%). Although the approach can be conducted without flexible components, a trade-off between the device strain limit and the component density is inevitable. The third strategy is a promising candidate for stretchable electronics, which fills the required conductor and semiconductor materials into elastomer support layers to impart device-level stretchability. For example, a capacitive-based touch sensor was proposed where a ZnS-doped elastomer and two LiCl ion-filled hydrogels served as the dielectric layer and electrodes, respectively. The skin sensor allowed a strain limit of more than 480% in the area. Although the all-stretchable-material strategy overcomes issues of tough surface and limited device density, the challenge of developing materials with high mechanical and electric properties still requires great efforts.

The aforementioned tactile sensing techniques are summarized in Table 1, together with their relative merits. In the table, “hysteresis error” and “response time” are about the temperature characteristics of thermoelectric sensors, and about the force characteristics of the other sensor types. The “cyclic stability” is represented by the maximum loading/unloading cycles for which the sensor maintains its sensing abilities.

Compared with the industrial needs for robotic tactile sensing, the current technologies can merely reach a limited subset. For example, sensors with triaxial force-sensing characteristics encounter difficulties in high spatial resolution and robustness performance. Therefore, the mechanism selection is necessary to be weighed against the features of specific application scenarios. For instance, the sensor response time is expected to be lower than 10 ms in dynamic operations; therefore, the suitable techniques may include piezoresistive (3.1 ms), piezoelectric (0.01 ms), and optical (0.1 ms) methods. The needs for dexterous object manipulation are high-sensitivity (<5 mN) and multiaxis sensing capability, which are consistent with the characteristics of capacitive (0.2 mN), piezoelectric (0.2 mN), piezoresistive (1.5 mN), and electromagnetic induction (0.125 mN) based solutions. In addition, long-term and wide-range force detection applications place constraints on the sensor robustness; the appropriate structures are currently based on piezoelectric and piezoresistive techniques, which can withstand 10 000–375 000 and 4000–29 000 times cyclic forces, respectively. Stable object grasping and human–robot collaboration tasks present demands for flexible robot skin surfaces, the options to which are twofold: the first is to cover the skin sensors (such as piezoelectric and piezoresistive sensors) with elastomer layers, which allows stable electrical wiring and high sensor durability; the second utilizes sensors that have elastomer sensing layers (such as capacitive, inductive, and optical-based sensors), without adding other materials. Note that the elastomer layer will result in response slugging effects (e.g., the response time of capacitive force sensors is >30 ms).

2.2. Application

2.2.1. Object Recognition

In object recognition tasks, tactile sensing can powerfully learn the details of objects through the perception and analysis of pressure. Broadly detected information for object recognition includes hardness, friction coefficient, texture, and geometry. Subsequently in this section, we first explain how this basic information is obtained from tactile sensing, and then discuss its use in object recognition.

Hardness: To obtain hardness information, the boundary shape of the contact region between the tactile sensor and the object is analyzed. As shown in Figure 3a, for hard objects, only small distortion occurs at the contact location, indicating that the boundary shape is apparent. In contrast, the boundary shape is vague and of a large area when soft materials are pressed. By using this phenomenon, researchers first used a GelSight tactile sensor (soft-touch interfaces) to test the relationship between the force, strain, and hardness of the object with a fitting method, and then the two hardness estimations were linearly weighted to form a hardness prediction. The hardness detection accuracy rate was 97.82% and 94.94% for items with fixed shapes and arbitrary shapes, respectively. Yuan et al. used a recurrent neural network to model the changes in gel deformation over time, trained the neural network based on the tactile image sequence of the known hardness samples during the pressing process, and estimated the hardness of the unknown shape samples with an accuracy rate of 81.8%.

Friction Coefficient: The friction coefficient of an object is calculated mainly based on the force or deformation measured during the sliding process. For example, Obinata et al. divided the contact area when people grasp the object into a central fixed area and the surrounding initial sliding area. The latter slides first as the grasping force decreases. Gradually changing the size of the force to produce different degrees of initial sliding, the friction coefficient can be calculated according to the ratio of the two areas.

Texture: To sense the surface texture of objects, two features are calculated from a sliding touch: the frequency distribution and the average force. The former indicates detailed surface patterns, whereas the latter describes the overall frictional properties. Jamali and Sammut used the average force value and the main time-domain frequency components to classify the surface texture of living goods, such as a carpet, dishwashing sponge, and tile. With a support vector machine (SVM) model, six kinds of textures were successfully recognized with an accuracy of 97%. In Ward-Cherrier et al., the frequency in both the time and space domains was extracted as features for texture classification. Through a K-nearest neighbor (KNN)-based algorithm, the accuracy was 98.3% in 11 3D-printed artificial textures and 92.8% in 20 natural textures.

Object Classification Based on Tactile Features: Object classification tasks can be divided into two categories: 1) objects of similar shapes but different materials and 2) objects of different shapes. Tactile sensing can be used to recognize both of them but performs better in the first category. The reason for this phenomenon is explained subsequently.

Category 1: based on the calculation of the aforementioned tactile features, tactile-based techniques are capable of recognizing objects with different materials. For example, Hoelscher et al. used a tactile-sensing robotic finger that performed static contact and dynamic sliding at a fixed speed to extract features such as hardness, surface texture, and pressure (including mean value, variance, skew, and kurtosis). The random forest (RF) and SVM algorithms were implemented for classification, and the recognition accuracy reached 97.6% for 49 cuboid objects with similar dimensions. In Cao et al., tactile sensing arrays collected sequence images under
Table 1. Summary of the commonly used tactile sensing techniques.

| Mechanisms                           | Physical quantity       | Key materials                                                                 | Lower limit of detection range | Upper limit of detection range | Hysteresis error | Response time (ms) | Cyclic stability (s) | Advantages                                           | Disadvantages                                                  |
|--------------------------------------|-------------------------|-------------------------------------------------------------------------------|--------------------------------|--------------------------------|------------------|-------------------|-------------------|------------------------------------------------------|---------------------------------------------------------------|
| Resistive (piezoresistive)           | Triaxial force, normal force, humidity | PEDOT:PSS film, monocrystalline silicon, conductive CNTs, polymer composites, EMIES ionic liquid | Normal force: 0.0015–0.05 N  Shear force: 0.3–20 N | Normal force: 0.05–0.15 N  Shear force: 0.6–10 N | 7.5–25%          | 3.1–100 ms       | 4000–29 000       | Simple structure and fabrication process, low cost, good mechanical durability, suitable for large area | High power consumption, suffering heating issue, low detection sensitivity, high hysteresis |
| EIT [193–195]                        | Normal force            | Same with resistive/piezoresistive sensors                                   | Normal force: 0.1–1 N          | Normal force: 10.30 N          | 20–34.5%         | ≈10 ms           | /                 | Simple structure, thin thickness, suitable for large area | Low detection accuracy, high power consumption, low spatial resolution, sensitive to EMI noise, low temporal frequency |
| Capacitive [199–201]                 | Triaxial force, normal force, humidity, proximity | PDMS film, Al2O3 film, SiO2 film, TiO2 film, PLA film, polyimide film | Normal force: 0.0002–0.01 N  Shear force: 0.015–0.02 N | Normal force: 5–14 N          | 7–35%            | 30–100 ms        | 1000–10 000       | Multidimensional sensing ability, simple structure, suitable for large area, high detection sensitivity with micro-structures | Poor mechanical durability (with elastomer dielectric), sensitive to EMI noise |
| Electromagnetic induction [205–208]  | Triaxial force          | Magnetic iron, copper/aluminum coil                                         | Normal force: 0.000125–0.004 N  Shear force: 0.003–0.01 N | Normal force: 1.2–13 N          | 2.6–12%          | 62.5–111 ms      | 1000–20 000       | High detection accuracy, low hysteresis              | Poor mechanical durability, complicated structure, sensitive to EMI noise |
| Optics [133, 210, 211]               | Normal force            | Silica glass optical fiber, CdS/PbS photosensor                              | Normal force: 0.05–0.2 N       | Normal force: 0.3–10 N         | 6.3–20%          | 0.1–70 ms        | 1000–10 000       | Immune to EMI noise and temperature change, rapid response | Complex and fragile architecture, low spatial resolution, high power consumption |
| Piezoelectric [57, 212–219]          | Triaxial force          | PZT ceramic, ZnO nanowire, PVDF film, PVDF-TrFE film                        | Normal force: 0.0002–0.02 N  Shear force: 4.48–40 N | Normal force: 0.02–0.03 N  Shear force: 0.5–5 N | 3–5%            | 0.01–11 ms       | 10 000–375 000    | High detection sensitivity and output, Well suited for dynamic applications, low power consumption, simple structure | Cannot detect static force signal, suffering heating and humidity issue, inhomogeneous responsivity induced by boundary conditions |
| Thermoelectric [95, 220]             | Contact, temperature    | n- and p-type single-crystal silicon, p-SnTe/n-PbTe film, n-Bi2Te3/p-Sb2Te3 film | Temperature: ≈0.1 K  Temperature gradient: 90–125 K | Temperature: 0.35–2%  Temperature gradient: 600–2000 ms | 0.0002–0.001 N  Shear force: 0.3–10 N | 100–10 000       | 600–2000 ms       | Passive detection, low power consumption, low cost, simple and rugged in construction, support for wide temperature range | Low detection accuracy, reference is needed, low sensitivity and stability, single measurement dimension, |

a) EIT is a capacitive-based sensing technology whose readout process requires scanning the capacitance between the two electrodes in turn. Therefore, the response time mainly depends on the electrode number and readout speed of the backend circuit. The typical reading time is of the order of 10 ms.[22]
different operations (pressing, slipping, and twisting) of daily clothing. The raw data were input into a convolutional neural network (CNN) to extract spatial and temporal frequency features (the extraction results are shown in Figure 3b) and then predict the cloth’s material. After training, 100 pieces of clothing made of different materials, e.g., wool, cotton, and silk, were effectively classified with an accuracy of 80.20%. In Baishya and Bäuml, a humanoid robot performed sweeping motions on tubes from up to down to discriminate six tubes of the same geometry but different materials, including foam, paper, metal, wood, smooth plastic, and rough plastic (Figure 3c). Material features regarding friction coefficient and texture were calculated and fed into CNN-based classification algorithms, exhibiting a high recognition accuracy of 97.3%. Two fingertip-based sensors were used to classify round fruits using squeezing and desqueezing operations. The recorded time series data were processed by a KNN model, finally achieving an accuracy of 92.86% for seven kinds of fruits with similar shapes and different textures (e.g., grape, orange, and lime).

Category 2: for objects with distinct appearances (more common in life), tactile-sensing-based recognition techniques face challenges due to tactile sensors usually being only distributed in the palms of humanoid robots, or just on the fingertips, which are much smaller than the target object; therefore, the cognition of posture and appearance is limited. One solution is to record a series of dynamic touch processes to achieve a more detailed understanding of the local shape. For

Figure 3. a) The deformation images of a soft tactile-sensing interface when pressing on samples of different hardness. The samples in the first line have a lower harness (35), whereas in the second line they have a higher hardness of 72. Reproduced with permission. Copyright 2016, IEEE. b) The spatial and temporal features extracted from tactile data sequences in a cloth material recognition task. The highlighted regions are assigned with larger weights in the CNN network. Reproduced with permission. Copyright 2020, IEEE. c) The humanoid system proposed in Baishya and Bäuml that identified the tube-shaped objects’ materials through sweeping motions. Reproduced with permission. Copyright 2016, IEEE. d) The tactile-sensing system for object classification developed by Luo et al. along with the object samples, number of touching, and the collected tactile images. Reproduced with permission. Copyright 2015, IEEE. e) Schematic diagram of the two-fingered robotic tactile sensing system in Schneider et al. The robot grasped objects at different positions and then described each object with 5 features. Reproduced with permission.
example, Sundaram et al.\textsuperscript{[98]} used touch sensors distributed on the palm to collect data during the entire object-grasping process. Among the time series data, seven typical frames with greatly different features were selected through a clustering algorithm and then used for object recognition. An accuracy rate of around 92% was achieved for 26 household objects with different geometries, such as scissors, glass, battery, etc. Bhattacharjee et al.\textsuperscript{[99]} used the skin on the side of the robot arm to record the contact process with the object, extracting the maximum force, the size of the contact area, and the center of the mass motion vector as features. Using a KNN algorithm, the classification accuracy rate was 91.43% for the items of diverse sizes, such as a small medicine bottle and a large carton.

Another effective method is to perform multiple static sampling of objects at different positions, then rebuild the overall shape or directly extract features from each small image. For the former, the tactile sensor is supposed to move along a certain route to cover the whole object. For instance, Khasnobish et al.\textsuperscript{[99]} used an edge-tracking method to recognize the outline of regular-shaped objects, where the tactile sensor’s start points, moving direction, and moving distance were collected and described as chain codes. With the chain codes model, 12 classes of classical geometries (sphere, cone, cube, hemisphere, etc.) can be identified. However, the limited density of the sensors results in inadequate object shape rebuilding, especially in detecting 3D geometries. Therefore, features are obtained from each image, and the object is represented by the distribution of a series of feature vectors. For example, as shown in Figure 3d, Luo et al.\textsuperscript{[9]} used a small fingertip sensor to move and touch in sequence until the entire shape was captured; these small images were combined as descriptors, which were calculated by the L2 norm of the image edge direction histogram. With 15 touches on each object, 91.33% accuracy was achieved for 18 objects with sharp edges, e.g., comb, tweezers, plug, and wrench. Schneider et al.\textsuperscript{[100]} used two-finger sensors to capture multiple parts of each object, taking the centroid of the image as a descriptor (Figure 3e). After randomly grabbing ten times, the accuracy reached 84.6% for 21 simple-shaped industrial and household objects.

2.2.2. Human–Robot Interaction

In addition to interacting with inanimate objects, humanoid robots are also expected to interact with humans, which is called human–robotic interaction (HRI). HRI has a wide range of applications, including user emotion recognition, intention recognition, and human–robot collaboration.

Touch is a significant way for people to convey emotions.\textsuperscript{[100]} People can use touch to communicate with humanoid robots, treat them as partners, and treat diseases related to affective disorders. Cooney et al.\textsuperscript{[11]} used sensors all over the robot’s body to recognize people’s movements, including hugging, touching, checking, hitting, etc., and to analyze the positive, neutral, and negative meanings contained in the movements. Steihl and Breazeal\textsuperscript{[102]} attempted to distinguish between pleasant and painful touches of robot arms. Altun and MacNeal\textsuperscript{[102]} differentiated human emotions from fixed touch gestures, including nine such as sleepiness, relaxation, and excitement, though the recognition success rate was disappointing (the system is shown in Figure 4a). Apart from this, as shown in Figure 4b, touch interaction helps to treat tactile-relative diseases.\textsuperscript{[103]–105} For example, studies\textsuperscript{[104,106,107]} have developed humanoid robots to interact with those with autism, perceiving the patient’s touch patterns (strength and frequency). These robots are capable of providing feedback on changes in facial expressions and gestures. The results showed that this not only improves the efficiency of treatment, but also overcomes patients’ emotional and social barriers.

Touch can be used to interpret and predict a user’s intention. There are two main applications of touch in robotics: one is to use human touch as an instruction to directly control the movement of the robot; the other is to infer human intention based on the movement characteristics and then automatically adjust the robot’s behavior to complete human–machine collaboration. In Wu et al.\textsuperscript{[108]}, a spherical fixed handle collected the force distribution applied by users’ hands to classify 16 types of intentional actions and to switch between the actions. The results of the research provide reference values for conveying user instructions. In Liu and Hao,\textsuperscript{[109]} the user interacted with the robot through a linearly movable handle. The user’s touch force and movement speed information were measured, and the trained neural network predicted the user’s intentional speed in collaboration. YWang et al.\textsuperscript{[110]} studied the changes in the touch-force sequence perceived by the robot when the human and the robot operate the same object and the intention of the human in different moving directions. Takeda et al.\textsuperscript{[112]} designed a dancing robot based on touch (Figure 4c). By measuring the force and

\begin{figure}[h]
\centering
\includegraphics[width=	extwidth]{figure4}
\caption{a) Furry pet robot developed for human affect recognition.\textsuperscript{[102]} Reproduced with permission.\textsuperscript{[102]} Copyright 2015, Elsevier. b) A humanoid robot is utilized for the interaction with children with autism.\textsuperscript{[106]} Reproduced with permission.\textsuperscript{[106]} Copyright 2010, IEEE. c) Dance partner robot presented by Takeda et al.\textsuperscript{[112]} Reproduced with permission.\textsuperscript{[112]} Copyright 2007, IEEE.}
\end{figure}
moment of the contact part, it calculates the movement trajectory and estimates the next step of the dance, thus adjusting the step length to adapt to the movement of the human body.

3. Vision-Enabled Humanoids

Equipping humanoid robots with electronic cameras and range-finders can help them autonomously navigate and execute tasks in an unstructured environment. This section introduces the most commonly used visual sensors for humanoids and the applications of visual perception in different fields, as provided in Figure 5.

3.1. Sensing Mechanisms

The development of robot vision is inseparable from optical measurement sensor innovation. The vision sensors used in humanoid robots can be divided into 2D and 3D sensors. The former are represented by monocular cameras and can be divided into charge-coupled device (CCD) and complementary metal oxide semiconductor (CMOS)-based systems and can generate 2D images. In practice, it is generally used in image processing. The latter are based on various depth-sensing devices and can obtain 3D information around the sensor. Therefore, they are widely used in humanoid robot applications that require high environmental interaction. In this section, we focus on explaining depth-sensing technologies, including structured light, time of flight (ToF), and stereo vision.

3.1.1. Structured Light

The structured light method actively projects the coded optical patterns (e.g., sinusoidal fringe and speckle) to the object surface and collects and analyzes the reflected pattern, which is modulated by the object height. This method is highly accurate in indoor environments.

Figure 5. Synthesis of commonly used visual sensors and their applications in humanoids. Reproduced with permission. From the top-center picture, in the clockwise order: Stack item fetching, copyright 2016, IEEE. Grab a goblet at the nape, copyright 2011, IEEE. Material texture reconstruction, copyright 2020, IEEE. Mapping of a chess piece, copyright 2021, Can Stock Photo Inc. Object shape modeling by a TACTIP sensor, copyright 2017, IEEE. A cylinder robot arm based on vision sensors, copyright 2020, IEEE. Touch detection enabled by shadow images, copyright 2020, ACM. Secure operation area division, copyright 2012, IEEE. Pass items to robots, copyright 2021, Wiley. Humanoid robot for catch and throw games, copyright 2012, IEEE. Object recognition in occlusion situations, copyright 2019, IEEE. Segmentation and category identification of three objects, copyright 2018, Elsevier. Human face recognition, copyright 2020, Elsevier. Tool type identification, copyright 2019, IEEE. Tableware cleaning robot, copyright 2006, IEEE. A robot plays a balancing game on the tablet, copyright 2016, IEEE.
3.1.2. TOF

The ToF method actively emits light and measures the time delay or phase delay of the reflected light to calculate the depth, which also performs well in indoor environments.

3.1.3. Stereo Vision

The stereo vision method uses two cameras to capture the same object according to the triangulation method. The distance between each sampling point and the camera is calculated. This method is suitable for outdoor environments with ambient light.

The working principles, merits, drawbacks, and representative commercial products of monocular cameras and three types of depth-sensing cameras are described in Table 2. \[7,115\]

3.2. Applications

3.2.1. Object Recognition

Vision-based object recognition techniques are divided into two categories: the object’s global arrangement features and the local invariant features-based methods. \[116\] The former learns the global structure of objects by matching the object image with a reference image through cropping, zooming, or rotating. It requires the object image’s spatial distribution position and intensity to be similar to those of the reference image. This method is effective in the identification of specific items (e.g., face recognition\[117–120\]), but it struggles in situations where the object is partially occluded, deformed, or the sensor’s perspective is changed. The latter extracts typical local features in the image and uses scaled and rotation-invariant descriptors to represent and recognize objects. This type of recognition method is more robust and can work from different perspectives when the target object is partially occluded. Therefore, this method is suitable for object classification tasks with multiple shapes and different perspectives (e.g., the classification of commonly used living objects\[10,121–124\]).

In humanoid applications, based on these techniques, diverse objects have been successfully recognized ranging from living goods to industrial components. For example, in Mahmud et al.\[117\] a global matching method was used to recognize human faces. The authors used principal component analysis (PCA) to simplify high-dimensional images into more compact encodings of their appearance attributes and then projected them onto the feature subspace to match the features with the training images. The algorithm finally reached a 96% recognition rate based on the JAFFE face database. Yu et al.\[10\] developed a max-pooling convolutional neural network (MPCNN) for object recognition and achieved estimation by extracting spatial-invariant features. With a dataset of five objects (pen, cup, box, coke bottle, and screwdriver) and a total of 44 poses, 94.5% object recognition accuracy and a 5/C14 pose resolution were achieved. Maturana and Scherer\[121\] proposed a 3D CNN to classify 40 categories of furniture, e.g., sofa, bookshelf, and stool.

### Table 2. Comparison of 3D vision sensors.

|                      | 2D image sensing technology | Depth-sensing technology |
|----------------------|-----------------------------|--------------------------|
|                      | Monocular camera            | Structured light         | ToF                      | Stereo vision            |
| Active/Passive       | Uses convex lens to focus scenes on the camera’s image plane (maps 3D to 2D) | Projects coded optical patterns to objects and measures patterns’ variation | Directly measures the time delay or phase delay of the reflected light | RCB image feature point matching and indirect triangulation calculation |
| Working Principle    |                             |                          |                          |                          |
| Depth-sensing accuracy | /                          | 0.01–1 mm                | 1–10 cm                  | <1 mm                    |
| Depth-sensing range  | /                          | Within 5 m              | Within 10 m              | Within 20 m              |
| Frame rate           | 15–400 fps\[^a\]            | 30–60 fps               | >100 fps                 | 15–50 fps                |
| Software complexity  | Low                         | High                     | Low                      | High                     |
| Immunity to environment parameters | Affected by ambient light | Affected by ambient light; greatly affected by object reflection | Less affected by ambient light and object reflection | Affected by ambient light and object texture |
| Application          | Suitable for both indoor and outdoor applications; cannot be used at night | Suitable for indoor applications | Suitable for indoor applications | Suitable for outdoor applications; cannot be used at night |
| Commercial products  | Keyence IV 500, Cognex DVT, Panasonic PV Basler ace | HP 3D Scanner Pro, Microsoft Kinect V1, Intel RealSense | Intel SoftKinect DepthSense, Microsoft Kinect V2, Lenovo Phab, ADI ADSD3100 | Leap Motion, Stereolabs ZED, MYNT EYE |

\[^a\]fps: frames per second.
which were recorded from 12 different perspectives with a depth camera. The point could data were first segmented into small volumes and then represented by spatial occupancies. Afterward, the data were trained in a CNN to extract local features and match with labels, finally reaching a 92% recognition rate.

### 3.2.2. Object Manipulation

Apart from recognition, vision can assist humanoids in operating objects. There are two main technologies used in this application: motion planning, which calculates the optimal movement path and operation by establishing a 3D model of the object and the environment; and the integration of vision in feedback control to perform real-time adjustment of actions.

For the former, the supervised learning method is usually used for training the robot to recognize the method to grasp the object based on vision. For example, researchers used the probability model to determine an item’s optimal grasping area, such as the cup handle. In Le et al., the optimal contact points for grabbing were determined through an SVM model from the object’s point cloud data. In Jiang et al., an SVM method was used to extract the appropriate rectangle range for grabbing, and to provide the position, direction, and opening angle ranges of the holder (Figure 6a). A CNN model was used to analyze a top-level object’s effective position and angle in a stacking situation (Figure 6b).

The latter is often used for fine adjustments in fixed actions. For example, Claudio et al. established a visual servo system for grasping and handling objects with both hands, as shown in Figure 6c. Vision was used to track the target position and posture and evaluate the current robot hand and operation by establishing a 3D model of the object and the theoretical model for grasping. Song et al. combined planning and feedback control to develop a robot for grasping route planning in a multiobstacle environment. It uses a 3D camera to build object models in the scene and establishes a potential field model that considers the gradient distribution under the target object’s coexistence and the obstacle. As a result, the robot arm’s optimal position at the next moment is planned in real time.

However, vision-based physical operations generally experience problems with object occlusion. When grasping an object, the robot hand’s occlusion will deviate the grasping position from the expectation, affecting the next operation, making it difficult for humanoids to perform precision operations such as stacking objects and using gadgets.

### 3.2.3. Vision-Based Tactile Sensors

To perform diverse meticulous hand operations, humanoids’ tactile sensors are expected to have high resolution. However, current tactile sensors suffer from difficulties in obtaining high density due to the large data bandwidth and complex power wiring arrangement. In contrast, visual technology is suitable given its high resolution as well as simple wiring and manufacturing process, indicating the potential of precise tactile detection. It is worth noting that, unlike traditional optical sensing methods, which arrange individual transducers to detect touch at discrete positions, the vision-based techniques use cameras to capture the holistic deformation of one piece of sensing medium and use image processing software to locate and track touch events. The sensing medium is usually an object with great elasticity, such as a piece of rubber or a balloon, and the deformation quantity is measured by visual cues of preprinted patterns or markers.

The first prototype of this technology was described by Sato et al. in 2008. They embedded upper and lower levels of marker arrays in different colors in a transparent elastomer. A CCD camera was used to track the relative displacements of these two layers of marker points under the elastomer (Figure 7a), and elastomer theory was used to calculate the applied force vector field. However, the theoretical model’s poor accuracy in complex shape calculation (e.g., robots’ fingertips) limits its applications. In Obinata et al., the researchers use signal-layer markers and fitting methods instead to model the applied force vector on unclassical shapes. Dot arrays were embedded on the surface of a hemispherical transparent rubber, illuminated by an LED, and imaged by a CCD camera underneath. By means of fitting, relationships between the dots’ displacement and diverse tactile parameters were established: normal force was measured by light intensity, tangential force was measured by lateral displacement of markers, and the rotation angle and torque were measured by the displacement vector field of dot arrays. Similarly, in Lee et al. markers were embedded on

![Figure 6](image-url)
the surface of an elastic cuboid and cameras were used to record the variations in the locality and size of the markers (Figure 7b). When force is applied to the elastomer, the distance between the marker and camera shrinks and the area of markers expands. The Gauss fitting method was used for estimating the touch position and deformation. However, the responsivity of this sensor is relatively low and the position estimation error is large (1.675 mm), originating from the small size of the permitted deformation.

To further enhance the responsivity, Ward-Cherrier et al.[133,134] developed a sensor named TACTIP to emulate the form of tactile corpuscles in human skin. As shown in Figure 7c, they placed thin needles in the elastic rubber shell and tracked the displacement field of those needles by an inside camera. Compared with markers, the thin needles significantly enlarged the gradient of displacement, which improved the location precision to 0.2 mm, far beyond the resolution of human skin. The GelSight sensor[135] improved the lighting of cameras using three colors of light (red, green, and blue) from different directions to illuminate the elastomer and using cameras to record the RGB images (Figure 7d). Based on the look-up table method, the normal field of every marker was calculated from the intensity of three kinds of light (Figure 7e,f), which increased the accuracy of the estimations of pressure, strain, and shape.

Table 3. Typical vision-based tactile sensors in humanoids.

| Sensor       | TACTIP[133,134,189] | Fluid-type touchpad[222,223] | IASTS[132] | GelSight[81,82,135,185,224] | OmniTac[186] | VTacArm[187] |
|--------------|---------------------|-----------------------------|------------|-------------------------------|--------------|-------------|
| Structure    | Rubber skin with pins/LEDs/camera | Elastic membrane with dots/red-colored water/LEDs/camera | Fabric skin with dots/LEDs/cameras | Elastomer gel with triangles/multicolored LEDs/camera | Elastomer gel (no marker)/multicolored LEDs/camera | Elastomer gel with dots/LEDs/camera |
| Device usage | System: 40 × 40 × 85 | Touchpad: 20 × 20 × 13 | System: 153 × 165 × 48 | Touchpad: 23.4 × 22.6 × 2.0 | System: 33 × 33 × 35 | System: 50 × 50 × 175 |
| Sensor dimensions (mm) | Product: 40 × 40 × 85 | Prototype: 20 × 20 × 13 | Touchpad area: 133 × 91 | Touchpad: 23.4 × 22.6 × 2.0 | System: 33 × 33 × 35 | System: 50 × 50 × 175 |
| Number of markers | 127 | 408 | 154 | ≈330 | / | 250 |
| Marker space (mm) | 2.4 | 1.5 | ≈5 | 1.2 | / | ≈8 |
| Localization accuracy (mm) | 0.20 | 0.29 | 0.63 | 0.1 | 0.05 | 1.20 |
| Sensing functions | Three-axis force, touch position, slip, shape, vibration | Three-axis force, touch position, friction coefficient, shape, slip | Touch position, shape | Three-axis force, touch position, moment of force, shape, slip, hardness | Touch position, shape | Three-axis force, touch position |
Table 3 lists several typical vision-based tactile sensors and summarizes the structure, interval between markers, resolution of location detection, and calculation methods for tactile parameters. Compared with the skin-based tactile sensors demonstrated in Section 2.1, the application of visual techniques greatly enhances the spatial resolution to the level of human skin, indicating strong abilities in 3D shape perception and delicate robot operations. However, the sensor area is often limited due to the restrictions on image processing complexity and camera shooting distance; thus, such sensors are placed at robots’ hands and fingertips, rather than a large area of the whole body.

3.2.4. HRI

Using visual sensors to track targets can achieve a variety of HRCs. For example, Kober et al. developed a humanoid that pitches and catches a ball with humans, as shown in Figure 8a. The robot uses vision to locate the ball position and a Kalman filter to predict the capture position and time. Then the robot’s hands and joints are calibrated to the visual coordinate system. The robot hands can be positioned to the predicted point. A quick catch and throw cycle were verified in experiments. Melchiorre et al. developed a human–machine handover control strategy. The visual sensor obtains the operator’s motion, and the path planning algorithm drives the robot to move to the operator’s hand. The handover object posture is adjusted to imitate the operator. In Wang et al., a human–robot cooperation system for minimally invasive surgery was proposed. The robot holds a camera (in an endoscope) and uses the perspective projection method to track the medical device’s position to assist the surgeon.

To achieve safe HRI, Rybski et al. proposed a 3D camera-based collision avoidance method using 3D camera data to identify and locate foreign obstacles in the human–robot environment and update the safe zone and danger zone scopes in real time according to the position and movement of the robot, as shown in Figure 8b. In another study, the operator’s hand was recognized and tracked, its closest distance to the robot was calculated, and the appropriate safety strategy was accordingly selected in the following four tasks: warning the operator, stopping the robot, moving the robot, and keeping away from the operator.

According to the previous sections, the differences between vision and tactility are summarized in Table 4. Vision has a wide field of view and can collect the target’s global data, which is conducive to overall shape analysis. Vision is noncontact, long-distance measurement that provides advantages in tasks such as forecasting and planning. Vision more easily acquires data and has a high image resolution. A typical depth camera, Kinect II, has a resolution of 512 × 424 pixels. However, some uncontrollable factors affect its performance, such as lighting, occlusion, and object posture, and it requires more computing resources. In contrast, tactile sensation focuses on the perception of local information. The sensor scale is usually smaller than the measured object and can only provide limited geometric details. It requires multiple or long-term contacts with the object, and the data collection process is complicated. The sensor pixels are typically only around a dozen, leading to low spatial resolution. It is less affected by the uncontrollable environment fluctuations, but more affected by changes in the target’s position and posture. The advantage of tactility is that it can measure a variety of detailed features such as surface texture and hardness. It requires contact conditions between the sensor and the object, which are conducive to precise positioning and control.

![Figure 8](image-url)
4. Vision- and Touch-Enabled Humanoids

Vision and touch have different dimensions, frequencies, and characteristics. The combination of both in humanoid robots can, first, complement the shortcomings of the other; second, be used to develop a robot sensory system with higher performance; third, achieve more accurate object detection and grip; and fourth, implement a variety of novel hand-eye collaboration applications. In this section, we explain how these four benefits can be obtained.

4.1. Vision–Touch Fusion Methods

Information collected from vision and tactile sensors must be matched and fused by appropriate strategies to form a cognitive tool that can be applied in different scenarios. We divide the strategies of vision–touch fusion into three categories: data-level fusion, feature-level fusion, and decision-level fusion. Table 5 summarizes the basic modes of these fusion strategies.

Data-level fusion directly splices or correlates the data retrieved from sensors and synthesizes the information into the same pattern, from which features are extracted as the basis for decision-making. For instance, during the grab operation,[146] data were obtained from two sources: tactile data representing the blocked area of the gripper and the unblocked area received by an electronic camera. These data were stitched into one image and then input into the neural network for feature extraction. In Smith et al.,[147] the depth camera created a rough image of the 3D shape of the object, in which details were covered by the shape information measured by the tactile sensor.

Feature-level fusion extracts a certain number of features from vision and tactile data, separately, and then fuses those features. This method is widely used for the avoidance of the information registration in different formats, and the features can be easily adapted to various machine learning models. Researchers[148] used a CNN to extract features from vision and tactile data through non-negative matrix factorization (NMF). They aligned the features in their common subspace and fused them with the objective of maximizing the common information and subsequently clustering the objects. In Calandra et al.[149] the vision data and tactile data were regarded as the input of the neural networks. After processing these data separately, the feature information was combined into a vector and input to the fully connected neural network for object classification. In Bohg et al.[150] a visual–tactile cooperative control framework was proposed, which processed these two kinds of features separately and added them into the feedback control loop to plan the operation together.

Decision-level fusion processes the vision and tactile information separately, then makes decisions based on extracted features, and eventually stacks or weights the decisions. For example, in an experiment on book extraction,[151] a vertical grasping strategy (position) was provided by vision, while the horizontal strategy (force) was determined by tactile information. In an experiment on brush operation,[22] the vision sensor calculated the error between the actual and expected handwriting, and planned the

| Table 5. Flow charts of vision–touch fusion methods. |
| --- |
| **Vision–tactile fusion methods** | **Examples** |
| Data-level fusion | [18,146,147,161,162] |
| Feature-level fusion | [6,19,147–150,164,166] |
| Decision-level fusion | [20,22,23,151–153,163,165,175,176] |
| Combined fusion | [5,154,158,160,172,174,177] |
next moving direction; touch was used to adjust the posture of the pen holder and the stroke force; and the two were superimposed to determine the motion strategy. In the control framework constructed by Nelson and Khosla and Prats et al., vision and touch made decisions independently and then were weighted and combined as negative feedback channels.

Several of the modes mentioned previously can also be combined. For instance, after making action decisions with vision, the grasping mode can be adjusted with the fusion features of vision and touch.[15]

4.2. Applications

Based on the fusion methodologies presented earlier, plenty of advanced object recognition and manipulation techniques have been reported, which can be mainly divided into two categories: performance enhancement of conventional tasks and touch- and vision-enabled new applications.

4.2.1. Performance Enhancement

Sense Enhancement: By establishing cross-mode association between vision and tactile measurement, the knowledge obtained from one modality can be transferred to the other, consequently improving or extending the sensory capabilities.[154,155]

Kroemer et al.[156] developed a vision-based texture recognition method for material classification. The ultimate goal of this study was to distinguish different materials using machine learning to process materials’ photos. However, the resolution of visual information was insufficient. Hence, tactile sensing was used at the initial stage of creating machine learning models. More specifically, the humanoid saw and touched the materials concurrently, and then the visual and tactile data passed through a weakly paired maximum covariance analysis (WMCA) model and then were projected into low-dimensional feature space, in which features were weakly paired. After training, the model was able to support vision sensors to work independently and the recognition accuracy of vision-based texture recognition was increased from 90.58% to 95.15%.

Takahashi et al.[157] proposed a CNN model that extracts tactile properties through vision images, where the input data were the 2D RGB images captured by cameras, and the output data were the triaxial force sequence measured by tactile sensors. By training the network, diverse tactile features were obtained in the middle hidden layers. The result qualitatively showed that the feature distribution in the hidden layer was related to the softness and friction of the materials, indicating the feasibility of transforming vision data into tactile properties. In addition, Li et al.[158] developed a conditional–adversarial model based technology of mutual transformation between vision and touch sensor images. The images were collected from a 2D web camera and a vision-based tactile sensor. Regarding the transformation from vision to touch, as shown in Figure 9a, the inputs of the network consisted of two parts: a series of tactile–vision matching images as reference and a series of visual images. The outputs were the predicted tactile images. To evaluate the authenticity of the output images (Figure 9b), human participants were surveyed using questionnaires to determine the similarity between output images and actual images. The result showed an average prediction accuracy of 90.37% based on 400 samples, indicating the feasibility of mutual transformation between different sensory modes, potentially helping humanoid robots to obtain enhanced perception in a variety of restricted environment scenes.

Object Recognition and Tracking: For humans, vision and touch functions provide complementary attributes. To demonstrate that the integration of these two could perform better in object recognition, a texture recognition experiment was conducted by Heller[159] in which subjects were asked to use vision, touch, and their combination to identify objects. The results showed that the recognition accuracies of using single modes were close to each other (about 70%), but when both modes were used, a high recognition accuracy of 82% was reached. A potential explanation for this finding is that vision establishes the spatial sense and tactile sensing is accurate at local exploration. With similar cooperation modes, in the area of humanoids, a variety of techniques have been proposed to enhance the modeling ability and recognition accuracy of complex-structured objects.

Local Detail Reconstruction: In Wang et al.[160] touch and 2D vision were applied collectively to perceive accurate 3D object shapes and to further classify 14 categories of daily necessities. Missions were arranged into two steps: first, CNN was used to predict the outline, depth, and surface normal of an object from 2D RGB images, after which 2.5D sketches of the object were synthesized; then, tactile sensing was used to sense the height distribution of different locations and the sketches were improved to a 3D shape model. The results revealed that the 3D shapes of ordinary objects could be constructed by about 10 random touch events; if the touch events were increased to 25, detail demonstration was enhanced effectively. Bjorkman et al.[161] further studied the tactile strategies for detail exploration. As shown in Figure 9c,d, they applied depth cameras to construct an incomplete 3D model from a fixed angle, used Gaussian process regression to calculate the uncertainties of every position estimation, and then applied tactile measurement to the region with the highest uncertainty. The process was repeated until the system produced a result. Based on the methods described previously, the required random tactile inputs decreased from 25 to 12.

Internal Space Exploration: In Allen,[162] tactile and 3D vision measurements were used collectively to recognize objects with concavities and holes, such as a mug, where the vision perception was blocked. First, vision was used to scan a 3D surface model of the object, and the surface area was divided into three categories according to level of detail. Then, tactile sensing was used to explore the areas with fewer details. For instance, to model an object with holes, vision was used to determine the holes’ positions, shape, and central axes, followed by tactile sensors extending into the holes along the central axis for shape construction. In Güler et al.[146] vision and tactile sensation were fused, and the category and volume of invisible substances in closed containers (Figure 9e) were identified through simple grabs and squeezes. Depth cameras located the position grabbed by hands recorded the deformation of the surrounding area, while tactile sensation was responsible for recording the pressure distribution of the contact area. The information from these two dimensions was superimposed directly (Figure 9f) and then processed by an
SVM-based algorithm; finally, 95% accuracy was achieved for the classification of five internal conditions (including empty, liquid, and particles of different sizes).

Vision-Impaired Scenarios: When a humanoid is performing operational tasks, the object is frequently blocked by machine hands, resulting in visual damage. In Corradi et al., recognition performance including a blocked condition was compared between single and double fusion modes. Applying the maximum likelihood estimation method, the recognition accuracies of single vision and touch were 50% and 85%, respectively (within ten touches), and the accuracy was improved to more than 90% by simply multiplying the two probabilities. Researchers in the past have used different fusion strategies to estimate the position and pose of the object blocked partly by robot hands, which achieved

---

Figure 9. a) The technology of mutual transformation between vision and touch proposed by Li et al., in which a conditional–adversarial model converted vision to touch. For touch to vision, the same framework is used, and the input and output modalities are switched. b) Examples of vision-to-touch prediction results. Reproduced with permission, Copyright 2019, IEEE. c) The 3D object model reconstruction method. Copyright 2013, IEEE. d) Output models against number of touches (none, 1, 4, 12, and 54 times), where 12 times was found to be sufficient to model and recognize objects. Reproduced with permission, Copyright 2014, IEEE. e) The containers and robot platform used in Güler et al. f) The process of extracting and combining visual and tactile data. Reproduced with permission, Copyright 2014, IEEE.
the accurate tracking of objects in the complex operations of humanoid robots.

In Honda et al.,[163] the operated object was marked by various patterns, and the spatial positions of the markers were tracked and calculated using depth cameras. After combing the data measured by touch, the pose of the object was estimated by the least-squares method, with estimation errors of 0.8° and 2.1° for the x and y positions, respectively. However, application scenarios are limited by the premarking method. To solve this problem, Hebert[164] replaced the artificial marks with the extracted visual features from the color and texture pattern of the object. A Kalman filter was used to estimate the continuous changes in visual features and tactile force. Then, Bayesian estimation was used to fuse the results of the two models to estimate the pose and spatial position of the object. The average angle error of three axes was 1.51° and the position error was 5.2 mm. Bimbo et al.[165] used partial point cloud data measured by depth cameras to build the tracking model. Then, position and triaxial force information obtained by the tactile sensor was corrected to calibrate the model. Based on this technique, the error of single vision was reduced from 8.58 to 2.66 cm, and the estimation performance was improved by 70%.

Object Manipulation: During the interaction between humanoids and the environment, vision can be used to quickly locate objects and plan operating routes, and tactile sensors can provide accurate compliance and contact force perception, thereby enabling a variety of applications with fine and flexible operations, mainly including the stable grasping of various fragile and complex-shaped objects. Based on this, delicate object manipulation can be achieved.

To allow humanoids to safely manipulate fragile and deformable objects, Calandra et al.[149] proposed an effective strategy for exploring and grasping unknown objects with a relatively small force based on an action-conditioned CNN network. After several attempts, the robotic hand (Figure 10a) collects tactile and visual images under different grasping positions and pressures. The system can constantly update the plan to achieve the goal of grasping with the minimum force. Compared with the single visual mode, the success rate increased from 76% to 94%, while the contact force was reduced to one-third of the original (18 vs 6 N). To improve object grasping stability, Calandra et al.[166] used a CNN network to extract features from tactile and vision information. This method predicts the probability of successfully grasping the object at different positions and chooses the optimal program. The success rate increased from 80% in the single visual mode to 94%. Guo et al.[167] introduced the grip stability parameter to the model, obtained by the weight of the force and strain information measured by a tactile sensor. Therefore, the state of the object changes from stable/unstable binarization to a continuous index, enabling the model to compare the performance of different grasping strategies in detail. In 93.2% of the gripping experiments, the object was guaranteed not to fall after vibration, shaking, or other operations. Others have studied the real-time action adjustment strategy during the grasping process. Based on force analysis, Rigi et al.[168] presented a numerical method for real-time detection of slip during operation. As the initial sliding begins, the local contact area steadily decreases. With changes in the contact area and force distribution, the sliding situation can be judged in real time, allowing the holding force to be adjusted in time to maintain stability. The results suggested that the sensor can detect the initial slippage of a variety of objects in an unstructured environment and effectively prevent false positives caused by vibration, with an accuracy rate of 70% and an average delay of 44.1 ms. Wang et al.[169] applied a recurrent neural network (RNN) with a memory function to extract slip features from a time series of visual and tactile images and to report the slip trend, which was capable of predicting and preventing about 84.6% of slip events.

For delicate manipulation, the fusion of visual and tactile senses has remarkably contributed to achieving accurate positioning and force control. For example, as shown in Figure 10b, a humanoid robot was designed to successfully pull a sliding-type cabinet door to a specified position with tactile and visual feedback.[153] At the beginning, vision was used to identify the position of the door handle and guide the action; when contact was established, touch and vision were combined into a negative feedback channel to control the movement. The results illustrated that the tactile sense guarantees favorable contact and allows the robot hand to adjust the magnitude and direction of the applied force to ensure the smooth movement of the door along the horizontal direction; however, when only controlled by vision, the hand and the handle were difficult to operate, and even disengaged. Schmid et al.[170] exploited a similar control technology in the implementation of pushing and pulling a drawer. When the robot hand touches the crossbar on

![Figure 10. a) The two-finger robotic hand for object grasping proposed in Calandra et al.,[149] as well as an example of the collected tactile data in an optimized grasping plan. Reproduced with permission.[149] Copyright 2018, IEEE. b) The experiment setup for a humanoid robot pushing open a sliding door, consisting of an external camera, a mobile humanoid, and tactile sensors installed at its fingertips. Reproduced with permission.[153] Copyright 2009, IEEE.](image-url)
the drawer, the tactile sense can judge the exact current relative position of the two to decide the appropriate method to tighten the holder. In the task of rotating a handle,\cite{171,172} the force and position information collected by the tactile and visual sensors was used to estimate the power at the current moment and project the torque at the next moment, as well as maintain the precise control of the rotation angle and speed. Kumar et al.\cite{172} developed a human–robot cooperation micromanipulation system based on force and visual feedback to insert a micropipette into animal cells to inject DNA. This process was carried out semiautomatically. A CCD camera and the tip force sensor were used to determine the relative location and the contact state of the cell, complete the puncture, and remove and insert the tip using automatic planning. Only the injection speed and residence time were controlled by the user, which not only improved the control accuracy of position and force, but also reduced the workload. In the experiment, even untrained operators successfully completed the task.

Due to the ability of visual–tactile fusion to recognize the position and posture of objects under occlusion, as demonstrated in the previous subsection, the task of grasping and using gadgets not only improved the control accuracy of position and force, but also reduced the workload. In the experiment, even untrained operators successfully completed the task.

A humanoid robot that efficiently cleans up garbage was demonstrated.\cite{176} It strategically removes items from a pile on a table and puts them in a trash can. First, it visually segments the objects in the pile and judges whether it can be captured.

4.2.2. Fancy Tasks

Based on the object recognition, tracking, and operation techniques presented in the proceeding content, humanoid robotics are able to perform complex tasks that were traditionally considered impossible. In this subsection, we introduce some works in recent studies.

In Furrer et al.,\cite{174} a humanoid robot piled up a balancing vertical tower with irregularly shaped rocks, as shown in Figure 12a. Vision was used to build a 3D model of the rocks and calculate the center of mass (CoM). The robot was controlled by both tactile and visual feedback to find the best placement. The stacking goal was set to: maximize the support polygon (the smallest polygon containing the indirect contact of the stone) and make the support polygon parallel to the horizontal plane as much as possible so that the CoM of the stone would be on a vertical line (Figure 12b). The experimental results showed that the robot successfully piled more than three stones in 72.7% of the cases.

Agravante et al.\cite{23,175} developed a strategy that allows humanoid robots and people to jointly carry a table while preventing objects from falling off the table. As the system framework shows in Figure 12c, the robot tracks the posture of objects on the desktop using vision, calculates the current tilt angle of the desktop, and uses tactile sense to perceive the partner’s intention, such as movement speed, target lift height, etc. This information is integrated into the visual–tactile feedback control system, successfully achieving stable control of regular square objects\cite{175} and spheres.\cite{23} This example can be applied to family life scenes, such as moving furniture.

A humanoid robot that efficiently cleans up garbage was demonstrated.\cite{176} It strategically removes items from a pile on a table and puts them in a trash can. First, it visually segments the objects in the pile and judges whether it can be captured.
according to the number of visible surfaces. For graspable objects, SVM is used to predict the appropriate operating surface and actions, including push, pull, and grasp along principal axis and grasp along secondary axis. Then, touch and vision are used to control the operation to apply small disturbances, determining whether the operation is safe; if not, the humanoid switches to a suboptimal operation. The experimental results suggested that when cleaning a pile of ten items, the number of actions required by this strategy was close to that of humans (19.1 vs 18.8). Compared with the random selection operation (35.4), the efficiency was nearly doubled.

Prats et al.\cite{151} describe a library service robot that finds and reliably manipulates needed books. The CCD camera recognizes the label on the spine of the book and guides the robot to find the desired book. Then, through the mixed control of vision and touch, the robot inserts the holder into the bookshelf and picks up the book while maintaining the movement direction perpendicular to the book to avoid affecting the surrounding books. Books with an average thickness of about 1 cm were removed within 1.5 s.

Feng et al.\cite{177} proposed a humanoid robot for estimating the CoM of a stick-like object with uneven mass distribution using...
vision and touch. The robot fixes a point (grabbing position) as the axis of rotation, observes the deflection of the stick after being subjected to gravity, and analyzes the distance between the CoM of the object and the axis of rotation. Vision is applied for recording the shape and the current grasping position and the tactile sensor is used for recording the dynamic changes in force and torque. With a long short-term memory (LSTM) network based model (shown in Figure 12d), the recommended strategy had a CoM prediction accuracy of 80.0% for five simply shaped objects (e.g., wooden hammer, LEGO model, screwdriver) and was able to rapidly judge the grasping method of complex objects.

Kudoha et al.[22] developed a humanoid painting robot to observe actual objects and draw its 2D outline with a paintbrush, as shown in Figure 12e,f. Before starting, vision is used to build object models, extract edge information, and then simplify, as a reference image. In the process of painting control, with visual and tactile feedback, the brush is controlled in three aspects: the contact force with the canvas, the tilt angle of the pen tip, and the lifting speed at the end of each stroke. Adopting this hybrid control strategy, the robot smoothly used four fingers to finely control the soft paintbrush, and successfully depicted the contours of apples and people (Figure 12g).

In Section 2–4, humanoids’ capabilities in object recognition, manipulation, and HRI achieved by different perception modalities (vision, tactile, and their fusion) are presented. The characteristics of these three modes in practical applications are summarized and compared in Table 6 and 7. These presented examples demonstrate humanoids’ great potential to facilitate our daily lives. However, they still face some challenges in actual applications, which will be discussed in the following section.

5. Challenges

Intelligent humanoids will collaborate with humans and provide services to us in many circumstances. However, their successful broad use has not yet been observed. Many factors can contribute to this achievement; in this section, to align with our article theme, we discuss these from the aspects of tactile and visual sensing.

Table 6. Comparison of object recognition methods.

| Application                                | Sensor                      | Algorithm | Applicable objects                                      | Recognition precision | Advantage and disadvantage                        |
|--------------------------------------------|-----------------------------|-----------|--------------------------------------------------------|-----------------------|----------------------------------------------------|
| Vision-based methods                       |                             |           |                                                        |                       |                                                    |
| Human face recognition^[17]                | RGB camera                  | PCA       | 10 Japanese females with diverse expressions          | 96%                   | Advantage: High precision in shape-based object recognition tasks |
| Household object and pose recognition^[10] | RGB camera                  | MPCNN     | 5 objects (pen, cup, box, coke bottle, screwdriver) with 44 poses | 94.5%                | Advantage: Good performance for objects with different materials or textures |
| Furniture category recognition^[21]       | RGB-D camera^[4]            | VoxNet (3D CNN) | 40 furniture categories (sofa, bookshelf, closes tool, etc.) | 92%                  | Advantage: Ability in detailed shape reconstruction; less restriction from scenes and objects |

Tactile-based methods                       |                             |           |                                                        |                       |                                                    |
| Texture recognition^[33]                   | Piezoresistive force sensor | RF and SVM | 49 cuboid objects with different surface textures     | 97.6%                 | Advantage: Good performance for objects with different materials or textures |
| Material recognition^[34]                 | Vision-based shape sensor   | Time-frequency analysis, CNN | 100 pieces of clothing with different materials (wool, cotton, silk, etc.) | 80.20% | Advantage: Ability in detailed shape reconstruction; less restriction from scenes and objects |
| Household object recognition^[9]          | Vision-based shape sensor   | Image edge extraction | 18 objects with sharp edges (comb, tweezers, plug, wrench, etc.) | 91.33% | Advantage: Good performance for objects with different materials or textures |

Vision-tactile enabled methods              |                             |           |                                                        |                       |                                                    |
| Recognition of invisible objects in containers^[146] | Depth camera and vision-based force sensor | SVM | 5 internal conditions (empty, liquid, solid with different particle sizes) | 95%                  | Advantage: Enhanced precision of object recognition and position tracking; ability of detailed shape reconstruction; less restriction from scenes and objects |
| Household object recognition^[20]        | RGB camera and vision-based shape sensor | SVM, Bayes estimation | 10 household objects (pen, book, cup, etc.) | ≈97% | Advantage: High computational complexity; difficulties in data/feature matching |
| Precise shape construction^[140]         | RGB camera and vision-based shape sensor | CNN | 14 household objects (bottle, helmet, camera, bag, bowl, etc.) | 5.5 chamfer distance (25 touches) | Advantage: Enhanced precision of object recognition and position tracking; ability of detailed shape reconstruction; less restriction from scenes and objects |

^[4]RGB-D camera: An RGB camera with depth-sensing function (D: depth).
5.1. Trade-Off between Large Area, High Density, and Multifunctionality

In an ideal case, tactile function is expected to be enabled among all body areas of intelligent humanoids. In this way, intelligent humanoids will be able to sense environmental information just as humans do. Therefore, three highly desired attributes for tactile sensors in intelligent humanoids are: large area, high density, and multifunctionality. However, concurrently achieving these three attributes is very challenging. For example, to obtain multifunctionality and high density, diverse sensors with different sensing abilities are required to be integrated, indicating that micro-electro-mechanical system (MEMS) technology is required. MEMS are usually expensive; it is unrealistic to install MEMS-based tactile sensors over a large area. Therefore, in practical use, tactile sensors are normally deployed for the robotic hands. The sensor density is far less than that of human skin, and in many cases only the force signal can be detected.\(^{[46,61,178]}\)

### Table 7. Comparison of approaches for object manipulation and HRI.

| Application | Sensor | Methods | Applicable objects | Successful rate | Advantage and disadvantage |
|-------------|--------|---------|--------------------|-----------------|----------------------------|
| Vision-based methods | | | | | |
| Play pitch-and-catch game with human\(^{[136]}\) | RGB-D camera | Kalman filtering | A ball of palm size | 75% | Advantage Perform well in object positioning and path planning |
| Grasp objects with complex shape\(^{[6]}\) | RGB-D camera | SVM | 9 household objects (martini glass, glove, etc.) | 87.9% | Disadvantage Difficult to manipulate flexible or fragile objects; suffering from object occlusion |
| Remove objects from the top of a stack\(^{[127]}\) | RGB-D camera | CNN | Random stack of up to 8 fruits (banana, orange, carambola, etc.) | 96.98% | |
| Tactile-based methods | | | | | |
| Ballroom dancing with human\(^{[12]}\) | Six axial force/ moment sensor | Hidden Markov model, ANN | 3 human subjects (their dance actions and moving trajectories) | 95.12% | Advantage A direct and rapid method for human emotion and intention detection |
| Cooperatively moving an object with human\(^{[109]}\) | Six axial force/ moment sensor | RBFNN | A linearly movable handle | 83.81% force reduction | |
| Autism therapy\(^{[107]}\) | Triaxial force sensor | Correlation analysis | A high-functioning autism sufferer | 25.7% improvement of therapeutic effectiveness | |
| Vision–tactile-enabled methods | | | | | |
| Push a sliding door\(^{[173]}\) | RGB-D camera, pressure sensor | Visual servo, impedance force control | A sliding-type cabinet door with a hook-shaped handle | Nearly 100% | Advantage Enhanced performances in delicate operations; better adaptability with different scenes and objects; support various advanced applications |
| Insert wedges into a groove \(^{[9]}\) | RGB camera, Six axial force/ moment sensor | CCN | 3 wedges of various shapes (triangle, circle, and semicircle) | 78.7% | |
| Center of mass estimation and stable object gripping \(^{[177]}\) | RGB-D camera, pressure sensor | LSTM | 5 simply-shaped objects (hammer, LEGO model, screwdriver, etc.) | 80.0% | |

5.2. Algorithm Transferability

As stated in Section 4, diverse machine learning algorithms have been applied for analyzing combined tactile and visual information. A shared issue for machine learning models is poor transferability, which means that the machine learning model generated using one machine could produce different performance when used for another machine, even for the two machines fabricated by the same process. This issue does not need to be urgently addressed now, as intelligent humanoids are not yet in large-scale production. However, this will become an inevitable problem once a production line is prepared.

5.3. User Privacy Leakage

Intelligent humanoids are expected to be involved in most aspects of users’ daily lives. Hence, a huge amount of users’ personal data will be retrieved by the machines. Current smart
devices perform corresponding actions mainly based on users’ voice input.\textsuperscript{179} In contrast, vision-based humanoids can directly see all details of a user’s personal life and working environment. This will pose potential risks regarding users’ information leakage, which may result in severe financial and social issues. Security-associated worries have been broadly discussed,\textsuperscript{180,181} but no perfect solution has been obtained.

6. Conclusions and Future Trends

Humans interact with the external environment primarily by touch and vision. Inspired by this, an increasing number of studies in tactile- and visual-sensing-based intelligent humanoids have been reported, and experimental results bring to successful completion the seemingly impossible tasks. In addition, humanoids can start to generate an appreciable degree of cognitive ability with the help of tactile and vision perception, implying an improved understanding of the circumstances, which benefits the several application scenarios requiring human–robot collaboration. Combining tactile and visual information to interact with the surrounding offers humanoids the ability not only to explore unknown objects, but also to understand more deeply how humans sense and learn the real world around them. Although the ultimate goal of automatically exploring new environments and making smart decisions is still out of reach for intelligent humanoids, the promising results summarized and discussed in this article indicate an exciting future.

In the foreseeable future, achieving higher pseudocognitive ability will be an important direction for the development of intelligent humanoids. The pseudocognitive abilities of intelligent humanoids strongly boost their adaptability in complex scenarios and enable them to explore new scenes automatically. This increases the efficiency of HRC and turns into reality scenarios previously described in movies, such as autonomic learning, flexible HRC, and understanding human beings, as listed subsequently.

6.1. Autonomic Learning

Autonomic learning has been a key capability that intelligent humanoids are expected to possess, which helps robots develop strategies by exploring environments independently, rather than pretraining a large-scale dataset. In this way, people could assign intelligent humanoids with various tasks at the lowest costs; moreover, the robots will be able to work independently in some dangerous and time-consuming jobs (e.g., space exploration).

To demonstrate this exciting trend, in Fazeli et al.\textsuperscript{[5]} researchers from the Massachusetts Institute of Technology (MIT) presented a hierarchical-learning-based technique to allow a tactile- and vision-based intelligent humanoid to perform tasks through self-learning. Through seeing and feeling objects, these intelligent humanoids can intuitively play Jenga, just like humans. In Burger et al.,\textsuperscript{[180]} a mobile robot was developed to conduct chemical researches fully autonomously. The robot successfully improved the hydrogen production efficiency from water by six times by selecting suitable photocatalysts mixtures.

6.2. Flexible HRC

After equipping with smart perceptual functions, humanoids start to be involved in human-centered environments, especially in the medical, manufacturing, and service industries. These scenes normally contain plenty of irregular objects and rapidly changing surroundings, creating big challenges to the reliability, flexibility, and compatibility of the robotic systems.

In current HRC circumstances, the robots passively play an assistive role—their actions are almost led and controlled by human collaborators. For instance, the teleoperated humanoid robots for surgery, such as the well-known Da Vinci system (Intuitive Surgical, USA), help surgeries by merely providing doctors with real-time visual images and giving haptic feedback through vibrations.\textsuperscript{[183]} This leads to low cooperation efficiency due to the unsatisfied tracking speed and the high raw-data redundancy. Therefore, stronger data parsing capabilities (such as object recognition, path programming, and human intention analysis) are highly required for future robots to understand and adapt to any changes in the collaboration tasks.

6.3. Understanding Ourselves

We have a long history to attempt to understand ourselves. We try to learn the fundamental reasons behind different human behaviors. Apart from conventional study methods, intelligent humanoids pave a new way. For example, in Atkeson et al.\textsuperscript{[184]} the researchers investigate how the human brain generates behavior, through inverse kinematics and trajectory formation. However, this area is still at an infant stage because the functionalities supported by intelligent humanoids are not rich yet. Nevertheless, with the rapid research progress in the area of humanoids, more interesting results are expected to give us a deeper explanation of human behaviors.

Scientists and artists are inspiring each other. By learning the latest research outcomes (even through reading newspapers), artists depict time beyond technologies, such as R2-D2 in Star Wars and Jarves in Iron Man. In the meantime, scientists are frequently reporting that their research outcomes bring the things once only imagined in science fiction into reality. Therefore, the authors believe that the future of the humanoid will be driven by both sides. As we stated, the aforementioned perspectives are actually near-future predictions based on current technologies. The far-future development is now wearing a mask. But no matter which direction the humanoids move toward, one thing we can assure is that tactile and vision fusion-based techniques will feature in the story.

Acknowledgements

The authors acknowledge the funding of the National Natural Science Foundation of China (Grant no. 61803017 and no. 61827802) and Beihang University (Grant no. KG12090401 and no. ZG2016S19C8).

Conflict of Interest

The authors declare no conflict of interest.
Shuo Gao received Ph.D. degree in electrical engineering from the University of Cambridge, UK, in 2018. From 2017 to 2018, he was a research associate with University College London, UK. He is currently an associate professor at Beihang University, China. His expertise area is human–machine interactive systems. He has over 100 publications, including books, peer-reviewed journals, flagship conferences, and patents. In terms of industrial experience, he worked as an optical fiber system engineer at Ciena Corporation, Canada, from 2012 to 2013 and a technique consultant at Cambridge Touch Technologies Inc., UK, from 2013 to 2017.
Yanning Dai is a master’s student at the School of Instrumentation Science and Optoelectronic Engineering, Beihang University, Beijing, China. Her research focus is flexible electronics for human–machine interaction and health monitoring.

Arokia Nathan is a leading pioneer in flexible electronics, display and sensor systems. Following his Ph.D. in 1988 from University of Alberta, Canada, he joined LSI Logic USA and subsequently ETH Zürich, Switzerland, before joining University of Waterloo, Canada. In 2006, he joined UCL as Sumitomo Chair of Nanotechnology. In 2011 he was a Chair of Photonic Systems and Displays, Cambridge University. He is currently a Bye-Fellow and Tutor at Darwin College. He has over 600 publications, 4 books, and 110 patents and four spin-off companies. He is a Fellow of IEEE (USA), IET (UK), and Chartered Engineer (UK).