GTI Enabling AI Vision with 5G Networks

White Paper

GTI Enabling AI Vision with 5G Networks

White Paper

V1

Version	V1
Deliverable Type	□ **Procedural Document** √ **Working Document**
Confidential Level	□ **Open to GTI Operator Members** □ **Open to GTI Partners** √ **Open to Public**
Program Name	
Working Group	Cloud Robot Program
Project Name	White paper
Source members	China Mobile, Softbank, CloudMinds, Huawei, Skymind
Support members	China Mobile, Softbank, CloudMinds, Huawei, Skymind
Editor	China Mobile, Softbank, CloudMinds, Huawei, Skymind
Last Edit Date	05-02-2018
Approval Date	DD-MM-2018

Confidentiality: The GTI documents may contain information that is confidential and access to the documents is restricted to the persons listed in the Confidential Level. This document may not be used, disclosed or reproduced, in whole or in part, without the prior written authorization of GTI, and those so authorized may only use this document for the purpose consistent with the authorization. GTI disclaims any liability for the accuracy or completeness or timeliness of the information contained in this document. The information contained in this document may be subject to change without prior notice.

Document History

Date	Meeting #	Revision Contents	Old	New

Executive Summary

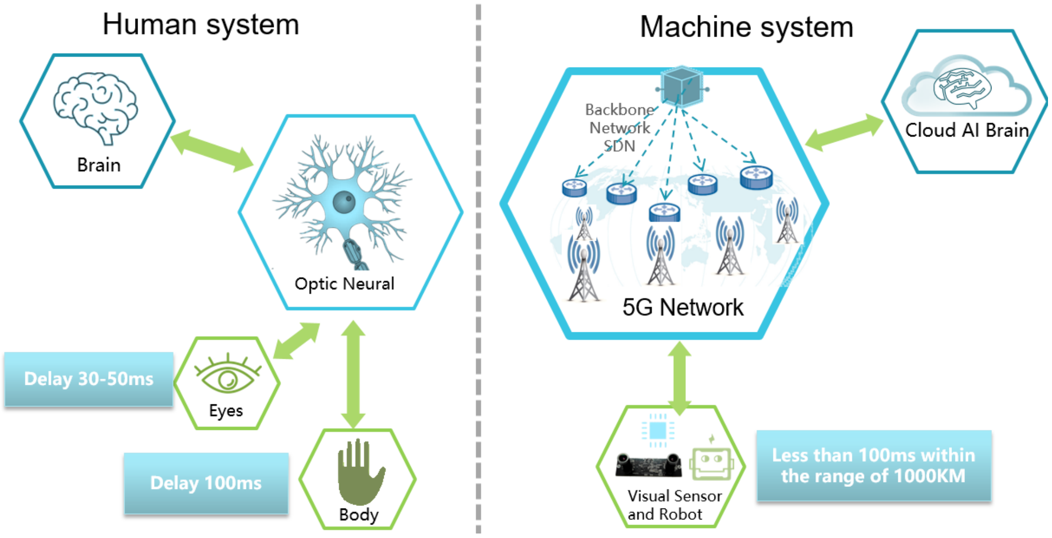
Table of Contents

1. Introduction

Computer vision based on artificial intelligence technology (hereinafter referred to as AI Vision) is the foundation of machine perception and cognition, and it is one of the fastest-growing areas of artificial intelligence. As reported, 91% of human information about the world comes from images captured by eyes. AI vision technology has the ultimate goal of making computers see and understand the world as humans.

To achieve this goal, machines must first have human-like "eyes" to see the world, including people, objects and the surrounding environment, and then transmit the visual signals through the optic network to the "brain" for signal processing. The signals then must pass through the brain's complex analysis process to make sense of the perceived object. The brain finally transfers the decision through the neural network to various parts of the body.

Figure 1. Cloud AI Vision System Based on 5G



As a corollary to how human see, process and react to their environment, in AI Vision, multiple visual sensors and deep learning chipsets will become the eyes for image capture, pre-processing and encode., 5G and backbone networks will provide the high-speed to transmit visual information, and provide feedback just as neural networks do. The vast amount of computational power in the cloud will be the brain for large scale parallel processing, and deep learning algorithms will act as the foundation for analysis, logic, reasoning, and decision making.

At present, there are many new trends in the field of AI Vision. One of the most prominent is the explosive growth of applications for this technology. These include use in robotics, public security, auto piloting, telemedicine, UAVs, and virtual reality among others.

The deployment of AI Vision requires an architecture that incorporates smart device, networks and cloud technology in mobile intelligent terminal devices, real-time response such as obstacle avoidance can be processed through terminal computing chips, intelligent decision making. On the other hand, requires large-scale computing and large data storage capabilities. That are often not supported by mobile terminals. Cloud -based artificial intelligence functions as the intelligent brain of the system. AI vision with real-time visual feedback demands fast information processing between the device and its "cloud brain", and requires wide network coverage, high speed, low latency and secure transmission. The introduction of 5G networks will address current communication challenges in AI Vision deployments.

2. Recent Advances on Intelligent Vision Technology

The 40 year history of computer vision has been marked by four major phases:

1) The Markov computational vision phase, in which computer vision became an independent discipline.
2) The initiative and purpose vision phase.
3) The multi-view geometric and layered 3D reconstruction phase.
4) The current deep-learning based computer vision phase.

In 2012, deep-learning based image recognition became widely regarded as the major breakthrough in vision technology, and has become the de facto standard for computer vision. The main advantages of deep learning based computer vision are:

- Deep learning algorithms are more versatile than the traditional algorithms, For example, faster RCNN, YOLO9000, SSD algorithms deliver superior results in facial, pedestrian and general object detection tasks.
- Transfer Learning: A widely used technique in deep learning allows the re-use of trained models on ImageNet for a variety of tasks including scene classification and object recognition.
- Data-driven feature design replaces human-guided feature design, which brings low cost engineering development, optimization and maintenance.

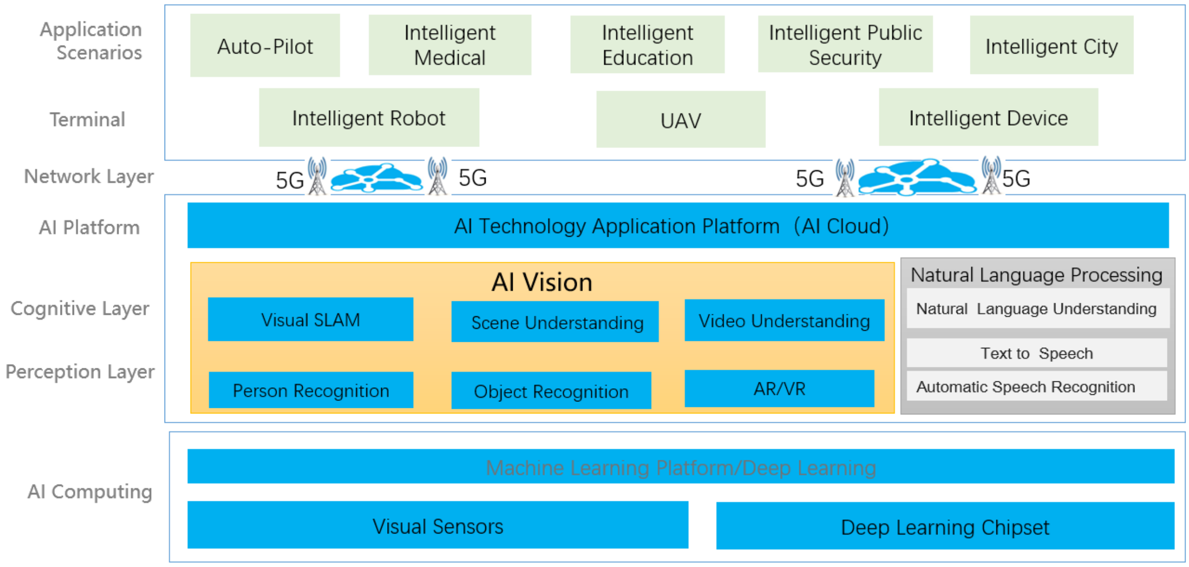
Successful applications of deep learning based computer vision include facial recognition, image search, object detection and tracking, all of which will enable fast deployment of computer vision in many vertical industries.

The following sections of this paper will analyze the development of the various major components of intelligent visual technology, including

1) Deep learning chipsets
2) Visual sensors
3) AI vision recognition of person, objects, and the environment
4) Visual SLAM
5) VR visual interaction
6) Intelligent robotic vision
7) AI vision Network architecture

The article also covers the cloud-based intelligent robot deployment architecture including smart devices, networks and the AI cloud.

Figure 2. Artificial Intelligence Architecture



2.1 Deep Learning Chipsets

Deep learning, by far the most dominant AI technique, relies on two major enabling technologies, enormous computational power and big data-based training.AI chipsets provide both.

With the increasing sophistication of deep neural networks in the field of computer vision(e.g. AlexNet proposed in 2012 to the ResNet proposed in 2015), the complexity between layers becomes a geometric multiple, driving the explosive growth in the demand for processor computing power.

Today, deep-learning based AI training is mainly conducted in the cloud, while inference can be carried out either in the cloud or on a terminal. Traditional CPU power cannot satisfy the computing power required by training and inference processes, therefore AI need new chipset architectures. FPGAs have been deployed widely in data centers.

Cloud computing: GPUs have taken a leadership role, but a wide variety of options coexist. At present, the GPU is leading cloud AI chipset market, ASICs such as TPUs are only used in the closed-loop ecosystem of a few industry giants.

Terminal computing: Limited by transmission latency and cloud security, AI Inference workloads are increasingly migrating to terminal devices. Terminal AI chipsets need to support large computing power, low latency, and have low power consumption. Typical industry verticals requiring such terminals are intelligent patrol and security surveillance, autonomous driving cars, cell phones / speakers / UAV / robot and other consumer terminals.

For today's AI uses cases, cloud AI computing and terminal AI computing will coexist for a certain period of time. To support various product features and functions, terminal AI will migrate to dedicated and customized chipsets, while cloud AI computing chipsets will become standardized and low cost for mass production. The future of the "AI brain" will still reside in the cloud.

2.2 Visual Sensors

There are many type of visual sensors, monocular, binocular (or multi-ocular), and depth sensors. Depth sensors include binocular stereo vision, radar detection, TOF technology and structured light-based depth cameras.

In the intelligent video surveillance industry, deep learning technology has revolutionized the ability of video analytics to accelerate the penetration of HD cameras into the field. The binocular camera and the panoramic camera enhance front-end data quality. With the combination of backend intelligence analytics systems, they will deliver unprecedented value, but at the same time challenge onto the video transmission network.

Intelligent robots, drones and autonomous vehicles, need to comprehend the visual signals gathered from complex surroundings and situations, including person, objects and environments, visual positioning, navigation and obstacle avoidance. Single vision sensors cannot complete such complex tasks, and therefore low-cost multi-sensor SLAM technology is becoming a standard in robotic devices. Vision-based path planning has also become a must-have capability.

In summary the integration of multiple visual sensors is a key requirement to support multi-sourced visual data. However, such AI vision capabilities to perceive, recognize, and understand all surroundings can only be made possible through the use of cloud technology.

2.3 Person Recognition

Recognition of person through AI Vision, includes facial recognition, portrait recognition and body posture behavior recognition. CNN-based deep learning in large-scale facial recognition retrieval scenarios, complex and diverse portrait recognition scenarios, or even gesture recognition scenarios usually resides either on a server or in the cloud.

- Facial Recognition

The deep learning method is rapidly replacing artificial feature selection for facial recognition. The latter technology includes face feature extraction, face verification or comparison, face recognition or face search, face attribute recognition, including age, gender, facial expression recognition, and face biometric detection are also an important areas of research.

In most face verification or comparison scenarios, the comparison can be done either at a terminal or in the backend cloud. **For face recognition against a large database of millions of facial images, face detection is generally implemented at the front-end. The face images are then sent to the cloud through the network, and the comparisons are done in the cloud.**

• Portrait Recognition

A variety of face information characteristics (e.g. gender, age, facial expressions, etc.), detection of human attributes (e.g. glasses, clothing features, backpacks, umbrellas, bicycles, etc.), can be used for rapid person recognition. The main application of video portraits is in the field of structural analysis. **Due to the complexity of portrait recognition, it is usually processed and analyzed in background by the cloud. The higher real-time analysis required, the larger cloud computing resources must be.**

• Body posture and behavior recognition

Human behavior recognition involves recognizing human behavior through temporal human gesture information. Such information includes single gesture estimation, multi-person gesture estimation, hand gesture estimation and other attributes. At present, the primary method is dynamic detection of the human skeleton. Such algorithm are very complex, therefore the accuracy from a single image still needs to be improved.

Occlusion has long been a challenge in gesture estimation. This problem can be addressed by use of the Generative Adversarial Nets (GAN). In actual application scenarios, dynamic detection of the human skeleton with multi-angle multi-cameras is used recognize human behavior posture. **Such data acquisition from 2D to 3D requires more computing power, therefore, the analysis is normally completed in the cloud.**

2.4 Object Recognition

Object recognition in the field of computer vision refers to the recognition of an object in a picture or an image from a set of video sequences. Object detection requires not only locating an object in an image, but also recognize all objects in the each frame of video.

Humans can easily recognize many objects in an image, although the objects in the image may be diverse. They may be different sizes rotated in different positions, or only partially visible. However, detection of thousands of objects in each frame of a video remains a challenge.

Object recognition needs to specify in advance the object category (sample base) that needs to be recognized. With this information, the objects appearing in a specific environment can be recognized and matched. For example, real-time detection/recognition can be done on a limited number of objects such as person and vehicles on outdoor roads. With complex environments, it is difficult to achieve real-time simultaneous detection/recognition of tens of thousands of objects. At present, intelligent analysis of video data structure in video surveillance is using object detection technology online or offline. **Such a process requires a very powerful GPU computing cluster or cloud computing for analysis and processing.**

2.5 Environment Scene Understanding

Environment scene understanding combines object recognition and scene information. Semantic image segmentation enables a machine to automatically divide the object area from the image and recognize the object category in each area, not only to recognize object but also the relationship between the objects. It also enables complex semantic comprehension, and scene mapping and recognition. In summary, environmental scene understanding requires a higher level of artificial intelligence and more computational power than do face recognition and object recognition.

Image caption, more complicated than object detection and recognition, use natural language to describe what happened in the image. For example, CVPR2017 research enables joint reasoning based on different parts of a descriptive video, and sorting of the text structure.

Task driven AI has always been the long-term goal in AI research field. This involves, humans using language to give commands to robots, and then robots using visual methods to observe the world, understand it and complete tasks.

Rich scene understanding, combined with language and task-driven computer vision are key development directions in field of computer vision. Scene understanding combined with language support further bridges the gap between human beings and computers, and task-driven computer vision also plays a critical role in the robotics arena, implicitly speaking, from visual images, provides access to more semantic information.

2.6 Visual Simultaneous Location and Mapping

Simultaneous Localization and Mapping (SLAM) are typically used in devices that require mobility. By acquiring data from sensors and through local or cloud computation, devices can autonomously generate their position and map of the surrounding. Visual SLAM technology is pivotal for robots and smart devices to interact by knowing self-location, surroundings, and how to autonomously move in the next step. It has a wide range of applications in areas such as autonomous driving, service robots, drones and AR. **In summary, it can be concluded that any smart mobility device possesses some form of visual SLAM capability.**

Human eyes are the main source of perception. Visual SLAM also has similar characteristics. It can acquire massive and redundant information from the environment, and possess superior scene recognition capabilities.

The computational requirements for laser SLAM are significantly lower than for visual SLAM. The Mainstream laser SLAM can be run in real-time on an ordinary ARM CPU, while basic visual SLAM requires a more robust desktop-level CPU or GPU support.

Future vision systems will include several functional modules such as positioning, map construction, motion planning, scene understanding and interactions, and all of which will bring more mobility to intelligent robots.

Multiple sensor fusion visual SLAM technology will give robots and devices an unprecedented degree of mobility. The visual SLAM relies on a richer three-dimensional environment map construction. **In the future, most of the mobile robots will rely on cloud based global positioning and navigation, as well as the shared environment visual SLAM map data and knowledge base required for multi-robots collaboration.**

2.7 VR Visual Interaction

Technology in 4K VR 360 video has matured, and 4K VR 360 video VR live and on-demand broadcasts are becoming more prevalent. However, the current viewing experience has not met user quality requirements, even in leading cloud platforms. The ideal bit rate for 4K VR is up to 30Mbps (with H.264 encoding).The minimum bandwidth required for seamless experience will reach 45 ~ 90Mbps and above.

VR 360 video quality issues are related to high network costs and insufficient bandwidth. Supporting VR 360 video bitrate requires advanced server -wide rendering and CDN distribution capabilities. Also, the higher the video bit rate, the higher the average rate of user access bandwidth required, and not all households have access to 100 Mbps bandwidth.

Insufficient bandwidth may impact user experience witch VR 360 video. Potential issues include frequent bounces, slow loading, and other problems. Video resolution, frame rate, bit depth, encoding protocols and projection methods all help determine minimum bandwidth requirements satisfactory performance.

2.8 Intelligent Robotic Vision

Intelligent robotic vision involves sensors that act as the "eyes" of the robot to enable it to achieve visual positioning and navigation for autonomous movement indoors and outdoors. Simultaneously, with the use of the camera's sensor, the robot can also recognize humans and objects in its environment. Such cloud enabled capabilities will provide robots with decision-making, analysis, logical reasoning.

At present, more and more robots use AI algorithms such as facial recognition, object recognition, positioning and navigation based on visual SLAM technology. By integrating these basic capabilities with natural semantic understanding, robots will have multi-modal human-like cognition levels.

AI vision research is focused on various recognition algorithms combined with environmental perception and understanding. From the deployment perspective, the smart terminal is used for visual information acquisition, and large-scale computing power to achieve perception, understanding and even decision-making. **Network coverage, transmission bandwidth and network latency are all critical facto**rs to enable intelligent robots to move around freely like humans.

2.9 Intelligent Robot Vision

Today’s AI mostly involves decision making, recognition, classification and prediction, whereas human cognition is based on logic, conversation, thinking, understanding and communication.

In the competition organized by ImageNet in 2015, the championship team, using deep learning algorithms, was able to reduce error rates to levels below those of the average human eye. However, in actual deployment scenarios, computer vision still cannot reach the acuity of human eyes due to environmental complexity, blurred images, low resolution, occlusion, and visual acquisition angles.

For intelligent robots requiring mobility, computer vision is used to perceive the environment, recognize person, understand semantics, and recognize objects, all while employing real-time visual feedback. Such perception and cognition typically require human intervention.

From the perspective of the data layer, visual data needs to be analyzed and processed. Three-dimensional environment map data and semantic knowledge base are needed to achieve the vision-based intelligent cognitive level.

In addition, from security perspective, visual information is transmitted from the visual sensor to the cloud, an artificial intelligence algorithm is calculated in the cloud to form a feedback loop to the smart terminal side, and end-to-end security protection is required.

The "brains" of intelligent robots must possess massive and complex computational power and therefore inevitably be located in cloud data centers. With efficient communication networks, the cloud “brain” will enable robots with visual perception, natural language dialogue and real-time control.

In summary, for AI vision to be used in general applications and implemented at the industrial level, computer vision must adopt a cloud architecture.

3. Cloud AI Vision Use Cases

Powered by a deep learning framework, computer vision is widely accepted in applications such as face recognition, video surveillance of public security, smart cities and intelligent robots.

3.1 Facial recognition Use Cases

Facial recognition is currently one of the most widely used artificial intelligence functions. It also generates great commercial value in many fields. The main types of facial recognition technology in use today include face detection, face verification and comparison, face search comparison and face attribute inspection (gender /age/expression, etc.).

Applications include but are not limited to smart buildings, smart enterprises, smart retail, smart communities, and smart conferencing. Specific scenarios include:

- Facial recognition for secure access: access control, attendance logging, concierge, visitor management, unknown visitor warning, and conference room management.
- Enterprise smart buildings and communities: visitor management, vehicle information and parking management, as well as video surveillance analysis.
- Video Surveillance: Personnel identification, behavior analysis, trajectory analysis, scene and environment analysis.
- Smart Retail: member identification, push advertising, visitor traffic and attribute analysis, online and offline data correlation and precision marketing.

3.2 Cloud Intelligent Robot

Most robots on the market today, still cannot replicate human intelligence levels. This is due to the limitation in the development of robot controller hardware the fact that artificial intelligence technology (including visual algorithms and natural language processing) has not yet reached the level of human intelligence. Robots can therefore perform only certain tasks under certain circumstances.

The improvement of robotic intelligence will depend on our ability to train with large amounts of data. Even with this advancement, there still exist the problems of data dimension disaster, and the inability of robots to derive knowledge from previous experiences.

With a controlled amount of supervised learning conducted by humans in the cloud, combined with continuous deep learning with large-scale computers, reinforcement learning based on human intelligence framework, and inferences from similar intelligence tasks, a continuous, iterative self-learning intelligent framework can be formed. When the robot encounters unknown problems, it will be necessary for humans to intervene directly to minimize risk and improve control in AI decision making.

Using human intelligence to augment artificial intelligence and provide error-free services is the path to the future in the development of cloud-based intelligent robots. The world's first cloud-based intelligent robotic company CloudMinds has been on the way. Such man-machine integration will be applied in more and more intelligent robot scenarios.

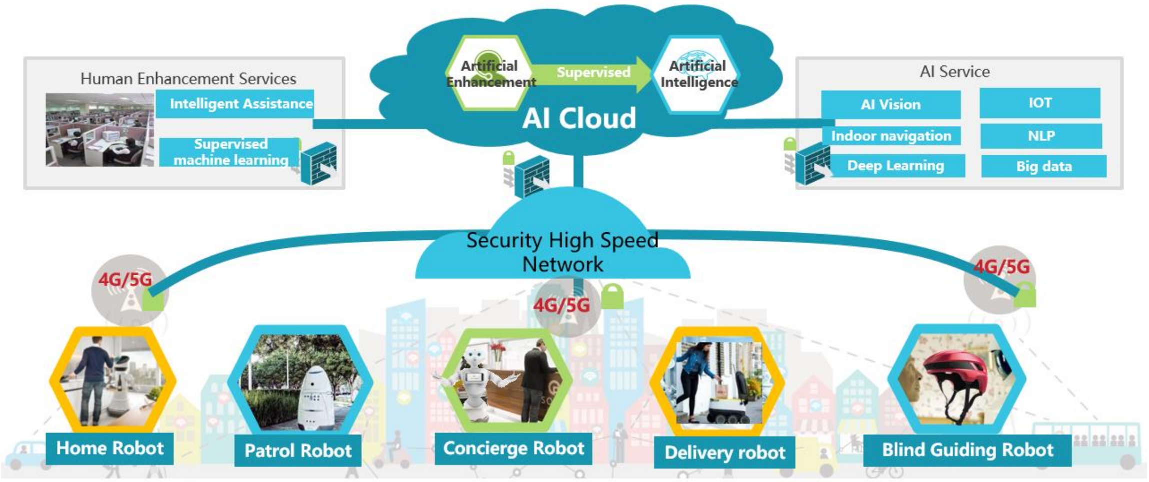


Figure 3. Human Enhanced AI Cloud and Intelligent Robots

To perform intelligent services, a robot needs to possess the ability to "listen, speak, see and move". In addition, these capabilities cannot exist in isolation. Instead, the intelligent cognitive system needs to be placed in the cloud, while sensors, and other hardware drivers are placed on the robotic body, with the two connected through a secure high-speed communications network.

3.2.1 Concierge Robots

Concierge reception robots are used mainly in airport lobbies, large commercial venues, shopping malls, bank VIP business lobbies and other settings to provide customized services to clients. Main application scenarios for concierge robots include:

1) Reception Robots: These robots use natural language processing technologies such as speech recognition, semantic understanding and speech synthesis to enable them to communicate with humans.

2) Business consulting and management: Tapping into various business knowledge databases, these robots answer business questions and process specific tasks to support day-to-day business operations.

3) Product Marketing: These robots effectively attract customers to new products, display product information through their local screens, and can also conduct marketing promotional activities.

4) VIP Customer Service: These robots specialize in identifying and greeting VIP customers based on face recognition, conducting business and promoting new products using natural language, and leading VIP customers into offices to private client managers.

3.2.2 Patrol Robot

The goal of patrol robots is to provide security patrols 24 by 7 in the community as would human security personnel. During its community patrols the robots navigate along community paths to discover if there are any suspicious vehicles, objects or people present. The robots also inspect community public facilities such as street lamps, benches and conduct general environmental health inspections. These functions require the use of big data scene recognition, facial recognition, license plate recognition, behavior recognition and other technologies.

3.3 Public Security

A fundamental requirement of intelligent security monitoring is how to make sense of a mass of crime scene video surveillance data with limited human analysis capabilities. For example, when a crime is committed, an investigator may need to read 2400 hours or 100 days of video camera data at normal speed. Among other issues, it is hard for an investigator to focus for a long time without getting tired. This is an urgent problem for the security industry to solve.

Intelligence surveillance based on computer vision has four key development areas:

1) Person analysis (facial recognition, body feature extraction).

2) Vehicle analysis (vehicle identification and vehicle feature extraction).

3) Behavioral analysis (target tracking detection and abnormal behavior analysis).

4) Image analysis (video quality diagnostics, video summary analysis.

Current security monitoring methods prescribe a front-end and back-end combination model. The intelligent front-end is limited by hardware computing resources, and therefore can run only relatively simple algorithms that demand high real-time performance. This approach however, ensures that algorithm upgrades and operation and maintenance become complex. **Back-end intelligent analysis (such as intelligent analysis server) can usually be configured with enough hardware resources to run more complex algorithms that allow for a certain amount of latency, both of which will co-exist for the long term.**

In the future, with breakthroughs in network transmission speed, intelligent monitoring and surveillance systems cloud-based public security video surveillance will become mainstream.

3.4 Intelligent “City Brain”

More than 80% of the information processed by the human brain every day is acquired through the eyes. Therefore, to create an intelligent city, the "brain" cannot be separated from "eyes" Surveillance system and cameras that reside in open spaces and public security areas in the city, as well as hand-held smartphone or kiosks can serve as the “eyes” that provide necessary information to decision makers in the city’s “brain”. An example of how AI is deployed in urban governance, is Alibaba’s "city brain", which controls 128 street traffic lights in the city of Hangzhou. This deployment has reduced road congestion and made scheduling more efficient. The core system uses artificial intelligence to enable centralized real-time analysis of the entire city, automating the deployment of public resources and creating a city powered by AI.

3.5 UAVs

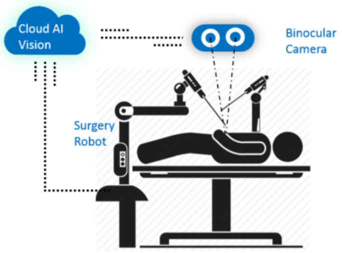
Because UAVs are low cost, highly flexible, secure, and not easily impacted by natural environment and terrain, they are increasingly used in infrastructure inspection, agricultural plant protection, logistics, monitoring, entertainment and rescue missions.

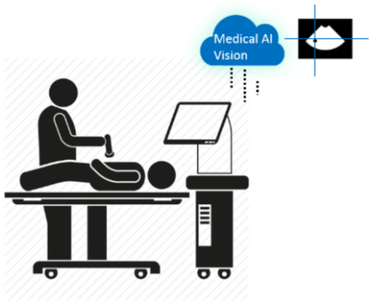
While conducting inspection operations, UAVs need to conduct intelligent surveys of infrastructure and perform analysis of grid inspection data they collect. This can include identifying deficient line spacing, anomalies in line amplitude and broken lines. In this capacity they replace highly paid inspection personnel. In the field of public security monitoring, UAVs can transmit real-time images, and use image recognition technology to search for specific personnel and determine the presence of dangerous goods at crowded public events. In natural disasters rescue missions, UAVs use visible light video and thermal imaging equipment to provide information regarding disaster-affected areas to command center in real time. UAV can also directly determine survivors' locations by using intelligent visual technology, increasing survival rates within the first “golden” 72 hour time window. In the field of entertainment, UAVs combined with AR/VR equipment can directly render and capture high-definition videos to stream processed video to AR/VR terminals. High-definition video streaming is very demanding on network transmission bandwidth, today’s wireless networks cannot support this premium user experience.

3.6 Telemedicine

Robotic surgery is a form of medical treatment that is based on robotics. It is comprised of positioning, sensing, and other related technologies to control an electronic robotic arm for the purposes of carrying out minimally invasive surgery. Before an automated or semi-automatic surgical operation, the robot must accurately locate the surgical position. Binocular vision and X-ray are two methods for robotic positioning. However, binocular vision positioning possesses considerable advantages over the use of X-rays. as it is non-contact, high accuracy, and non-radioactive.

Figure 4: Cloud AI Vision for a Surgery Robot





Medical image-assisted diagnosis is a medical procedure that incorporates artificial intelligence into the tagging of images for doctors to accurately assess a patient’s condition. During the examination, Ultrasound, X-ray, CT and MRI images will be transmitted to the cloud AI, and the AI will detect and tag any abnormalities. The tagged images will then be sent back to the doctor for further confirmation. For real-time medical applications such as ultrasound and endoscopy, real-time labeling of suspicious points can help doctors to complete an in-depth examination and significantly reduce misdiagnosis rates.

Figure 5: Medical AI Vision for Diagnostic

Robot guiding is an assistive service that uses artificial intelligence visual recognition technology to provide face recognition, object recognition, path planning and obstacle avoidance for visually impaired patients. The robot terminal (helmet or glasses, etc.) is equipped with a built-in voice interaction system and stereo vision camera. The camera transmits real-time video to the cloud. The cloud AI performs object recognition and converts information such as the name of the identified object and distance into a voice feedback to the user. The user is then instructed to take timely actions, such as bypassing obstacles, waiting for traffic lights, etc.

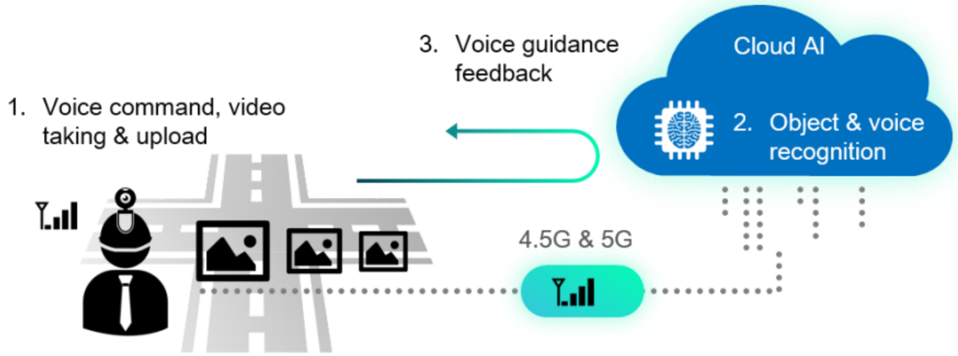

1. Voice command, video taking & upload
3. Voice guidance feedback
Cloud AI
2. Object & voice recognition
4.5G & 5G

Figure 6: Cloud AI Vision for Personal Assistant

4. Challenges in the Cloud Intelligence Vision

4.1 Intelligent Compute Level

It requires an enormous amount of computational power for today's deep learning based machine vision to achieve visual cognition. At the hardware level, AI cloud computing has created a competitive landscape in which GPU, TPU and FPGA solutions currently coexist. As a result, AI services cannot be flexibly migrated as traditional cloud computing services. This limitation is especially evident when large-scale distributed AI services are in operation. Reliability and usability often times present a large challenge. With the development of cloud AI computing and hardware standardization, these problems will be gradually resolved.

AI chipsets that support AI terminal computation capability have seen diversified developmental progress in smart security cameras, autonomous driving cars, mobile devices / speakers / UAV / robots and other consumer terminals. However, for industrial development of terminal AI products application-specific hardware development compatibility issues must be solved.

In addition, various deep learning frameworks have flourished simultaneously, including Google 's TensorFlow, Microsoft's CNTK, Pytorch, Caffe2, DeepLearning4j, Baidu's PaddlePaddle. Each framework has its own strengths in computer vision, natural language and application platforms. For intelligent robots that need "listen, speak, see and move" capabilities, the industry must utilize the superior capabilities of multiple deep learning frameworks described above to provide cloud-based AI services. Public cloud deployment is preferable to private deployment.

4.2 Visual Algorithm Level

With a large number of clean datasets, visual algorithms can design and train good models. However, in practical application scenarios, the image size, resolution, lighting, environmental complexity, degree of blurring, degree of occlusion, collection angle and other factors make it hard for artificial intelligence algorithms to solve common problems.

In an application scenario when a variety of visual sensors are used simultaneously to form a multi-source visual data fusion situation, it is critical to understand how to fuse the obtained data.

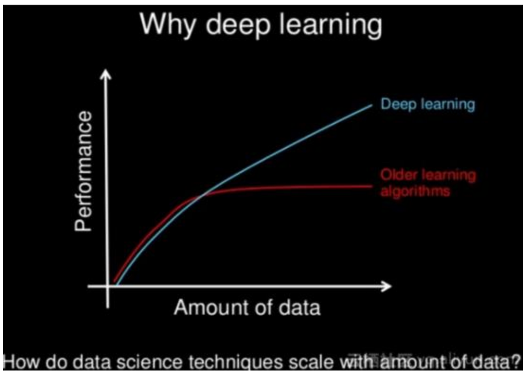
4.3 Data Level

Millions of images are typically needed to train an algorithm model to possess human-like vision capability. Therefore, the application of intelligent vision relies on big data sets. In the specific application scenario of AI Vision technology, it is particularly important to obtain the image big data in specific scenarios including:

1) In traditional industry verticals such as healthcare that have disaggregated data

2) With data is possessed by a handful of industry giants, and these companies do not have artificial intelligence research and development capabilities, and the start-ups with AI capability to mine the data do not have access to it.

3) In instance of data poisoning, in which although there are commercially available avenues to get data, there is data tempering during the circulation process.



4.4 Vision of the Intelligent Robot

Most robots on the market today process visual information using local processing. Their perception capabilities are limited to basic face and object recognition and sensory skills. They have yet to achieve higher level of intelligence such as environment cognition and memory. The main obstacle in the way is the large amount of processing power required for cognition. This power typically must reside in the cloud, and the low connection bandwidth and network delay to the cloud do not currently provide acceptable performance levels.

Future robots with cognitive level intelligence will thus necessity be based on cloud intelligence. However, **a key prerequisite is to solve the problem of the communication delays between the robot's hardware and its intelligent "Cloud Brain".**

4.5 Network Requirements

The path from robots to cloud includes video capture, encoding, network transmission, decoding, cloud AI processing, and audio warning.

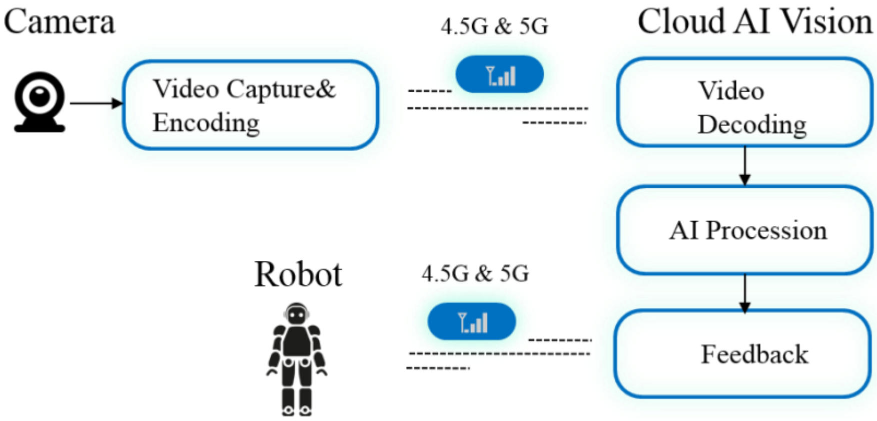


Figure 7: Cloud AI Vision link

In the AI vision system, the wireless network serves as a bridge between local devices and cloud intelligence. Network performance greatly affects the performance of the entire system. The following sections analyze network latency, bandwidth, security, and other characteristics of cloud-based intelligence.

4.5.1 Latency

The Latency requirement depends on the type of application. For public security application, 500 ms E2E latency is acceptable. For navigation of a robot with a low speed (< 6 km/h), 200 ms is sufficient. For personal AI assistant application such as for visually impaired guidance with sound, 100ms is the maximum latency.

Latency data based on AI cloud vision test (720p@36fps)

Process	Latency	Note
Video Capture	16-42ms	Including shutter, frame, H.264 encoding.
Network	20-50ms	Wireless network such as 4G.
Video Decoding	10-20ms	H.264 Decoding.
AI Processing	X00ms	Depends on the content and application.
Feedback	10-20ms	Command and result send back to robot.

4.5.2 Bandwidth

The resolution and bandwidth requirements of a single video are as follows:

Resolution	720P	1080P	4K	8K
Frame rate	30FPS	30FPS	30FPS	30FPS
Video Compression Standard	H.264	H.264	H.265	H.265
Bandwidth(Mbps)	1.5-2.5Mbps	4-8Mbps	20Mbps	50Mbps

Network bandwidth is recommended to be 1.2-1.3 times the required bandwidth when used in mobile network applications.

4.5.3 Security

The biggest threat to AI vision security is a loss of devices control (e.g. malicious attacks from hackers). However, there are a number of precautionary measures that can be implemented to ensure optimum security.

- At the network level, security can be implemented by isolation on the channel.
- AI vision authentication and related security mechanisms can be used for improved access security.

4.5.4 Coverage

Continuous and seamless network coverage is needed for AI vision systems that must operate over a wide area such as AGVs (Automated Guided Vehicles) and visually impaired guidance robots. Once the network connection is interrupted, the robot will immediately cease operations. Therefore, dependable and reliable network continuous coverage is a basic requirement of cloud-based AI vision applications.

5. The 5G Era-AI Vision

At present, 4G communication systems mainly provide a platform for interactions between people. In the future, 5G will not only provide people-to-people communications, but also provide communications between people and things, and things and things.

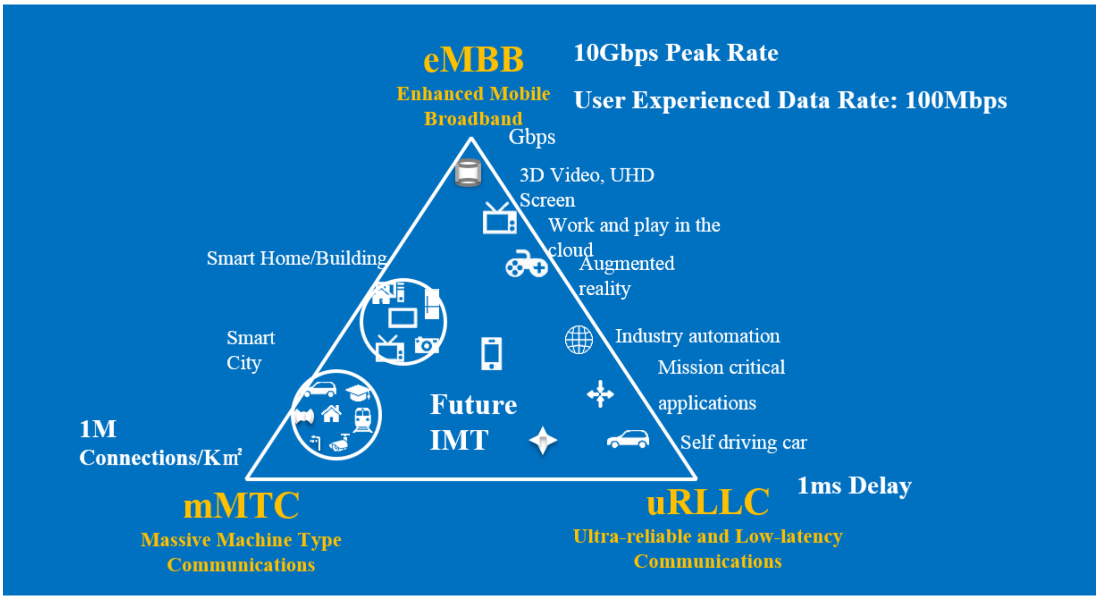

eMBB
Enhanced Mobile Broadband
10Gbps Peak Rate
User Experienced Data Rate: 100Mbps
Gbps
3D Video, UHD Screen
Work and play in the cloud
Augmented reality
Smart Home/Building
Smart City
Industry automation
Mission critical applications
Self driving car
Future IMT
1M Connections/Km²
mMTC
Massive Machine Type Communications
uRLLC
Ultra-reliable and Low-latency Communications
1ms Delay

Figure 8. 5G application scenario

The Cloud AI vision system uploads the video stream to the cloud and after it is processed downloads the result of the visual identification to an intelligent device. The Network requirements of the AI vision system must be supported both in 5G eMBB and uRLLC scenarios. 5G will be a key enabler for the cloud-based AI vision.

- Millisecond Level latency: Low latency is important for navigation applications, especially AI visual devices and high-speed robotics. 5G can support millisecond level end-to-end latency, whether it's delivering video to the cloud or sending cloud commands to devices and robots.
- Gbps Level Bandwidth: 5G uses a higher spectrum than previous networking technologies(4G,etc), which provides larger ,Gbps level spectrum bandwidth. A smooth viewing experience for 720p to 4K, 8K, AI (note: pixel resolution continues to increase) requires large uplink bandwidth. In addition, a number of AI visual applications involving 3D maps can also require large downlink bandwidth.
- Multi-level Coverage: 5G will support overlay networking and greatly enhanced multi-level coverage. AI vision devices can be connected with two or more radio links (In the event that one fails), the device’s connection will remain stable. 5G will provide high-quality connectivity for high-speed scenarios such as cars and trains.
- Network Slicing: 5G provides end-to-end network slicing for AI vision applications to support better QoS guarantees. At the same time, network slices are isolated from each other to provide more secure network services.

As a key component of the 5G solution, mobile edge computing (MEC) provides crucial support for AI vision.

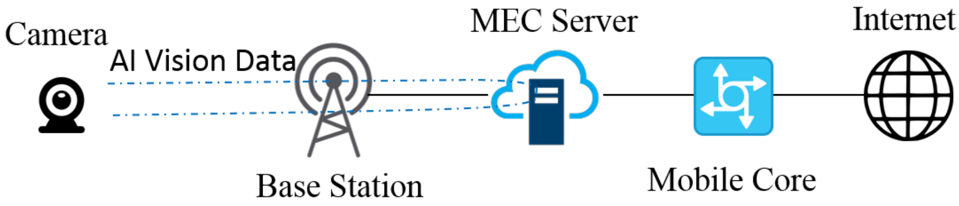


Figure 9. MEC for AI Vision

With MEC, operators can deploy content and services in closer to proximity to users and increase mobile network rates, reduce latency, and improve connection reliability. The user experience is significantly improved with the adoption of MEC.

MEC has the potential to support multiple functions: local content caching, artificial intelligence deployment, business optimization (based on wireless awareness), local content forwarding, and other networking capabilities. MEC is suitable for applications with high real-time requirements and large amounts of data such as the use of robots, AI applications, AR, V2V, corporate offices, MCDN, smart home, and IoT. MEC meets the system's requirements for multiple aspects of throughput, latency, network scalability, and intelligence.

6. Summary and Outlook

This paper summarize the development of artificial intelligence vision technology from the perspective its components, including vision chipsets, vision sensors, and neural networks built on deep learning algorithms. It also provides a comprehensive analysis of its practical applications in person, object, and environment recognition, visual positioning and navigation and so on. The future of AI vision will inevitably be based on cloud architectures and services delivered from the cloud for intelligent robots and devices.

Network related challenges faced today are centered on transmission bandwidth, delay, and security. The objective of the mobile 5G networks is to solve such issues as machine to machine communication. The advent of the 5G era will have a tremendous positive impact on the fields of AI and robotics.