UNITED ARAB EMIRATES
MINISTRY OF ECONOMY

الإمارات العربية المتحدة
وزارة الاقتصاد

(12)Patent Registration (B1)		**(12) تسجيل براءات الاختراع (B1)**
(19) United Arab Emirates		**(19) الإمارات العربية المتحدة**
(10) Publication Number	**AE P8326 B1**	**(10) رقم النشر**
(43) Publication Date	**2025-04-17**	**(43) تاريخ النشر**
(11) Patent Number	**8326**	**(11) رقم البراءة**
(24) Registration Date	**17-04-2025**	**(24) تاريخ التسجيل**
(54) Title of Invention		**(54) عنوان الاختراع**
Active Vision System for Achieving Sharp Vision		نظام رؤية فعال لتحقيق رؤية حادة
(51) IPC Classification	G03B 5/00 (2006.01)	**(51) تصنيف IPC**
(21) Application Number	**P6002246/2022**	**(21) رقم الطلب**
(22) Application Date	**31-10-2022**	**(22) تاريخ الطلب**
Filing Date:	**31-10-2022**	**تاريخ الإيداع**
(71) Applicant		**(71) مقدم الطلب**

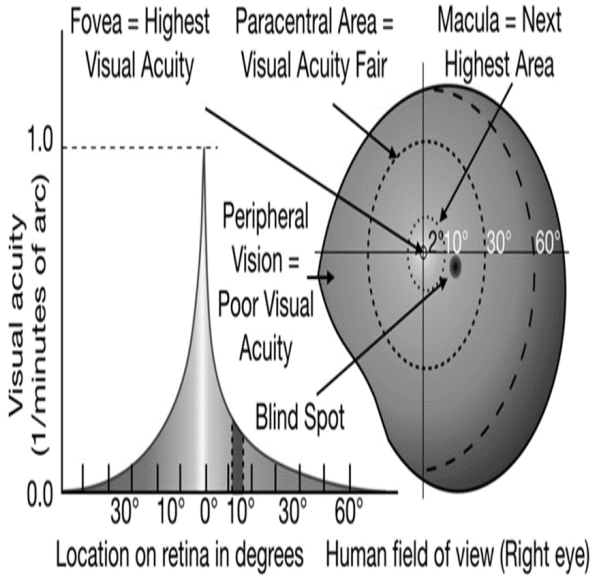
FOVEA TECHNOLOGY L.L.C, 44 Ad Doja 4 St, Musaffah, Abu Dhabi,
United Arab Emirates
,

(57) Abstract

Intelligent vision systems and methods which are capable of focusing on subtle visual cues at high resolution using principles of a foveal view coupled with an attention mechanism, that functions like the human eye are described. Intelligent vision systems and methods are based on artificial intelligence-deep learning neural network models that rely on a hierarchy of visual streams towards achieving human and greater levels of sharp vision. Intelligent visions systems and methods process multiple high-resolution field of view visual streams to create a foveal view representation of the environment which is fed forward to the neural network, which controls the pan/tilt positioning of the cameras based on an attention mechanism to achieve focus on a Subject of Interest (SOI).

(57) الملخص

يتعلق الاختراع الحالي بوصف أنظمة وطرق الرؤية الذكية القادرة على التركيز على الإشارات البصرية الدقيقة بدقة عالية باستخدام مبادئ الرؤية النُّقَيْرِيَّة المقترنة بآلية الانتباه، والتي تعمل مثل العين البشرية. تعتمد أنظمة وطرق الرؤية الذكية على نماذج الشبكات العصبية للتعلم العميق للذكاء الاصطناعي والتي تعتمد على تسلسل هرمي للتدفقات البصرية لتحقيق مستويات أعلى من الرؤية البشرية والرؤية الحادة. تعالج أنظمة وطرق الرؤية الذكية مجال عالي الدقة متعدد لرؤية التدفقات البصرية لإنشاء تمثيل رؤية نقري للبيئة التي تتم تغذيتها بإرسالها إلى الشبكة العصبية، والتي تتحكم في تموضع الدوران/الإمالة للكاميرات بناءً على آلية الانتباه لتحقيق التركيز على الهدف محل الاهتمام (SOI Subject of Interest).



(72) Inventor

Abdul Rahman Abdul Ghani Hafiz Abdul Ghani, 44 Ad Doja 4 St, Musaffah, Abu Dhabi, United Arab Emirates
Koganti Nishanth, 44 Ad Doja 4 St, Musaffah, Abu Dhabi, United Arab Emirates

(74) Agent

Dennemeyer and Associates S A

(72) المخترع

عبدالرحمن عبدالغني حافظ عبدالغني , 44 شارع الدجا 4، مصفح، أبوظبي، الامارات العربية المتحدة
كوجانتي نيشانس, 44 شارع الدجا 4، مصفح، أبوظبي، الامارات العربية المتحدة

(74) الممثل القانوني

دينيماير اند اسوسيتس اس ايه.

DESCRIPTION

Active Vision System for Achieving Sharp Vision

BACKGROUND OF THE INVENTION

Visual acuity is defined as sharpness or clarity of vision. Humans visually perceive their environment with very high levels of detail at very high resolutions (60 to 120 pixels/degree) in a 360-degree environment. This visual acuity allows humans to perform tasks such as driving and engaging in social interactions that require attention to subtle visual cues such as the intention of a pedestrian or facial expression.

Visual acuity is a function of the fovea which is the portion of the eye that has the capability to achieve the greatest acuity. At the center of the fovea, visual acuity is at its greatest. Moving away from the fovea, visual acuity progressively diminishes, thus establishing a visual representation model that is a hierarchical layering of varying degrees of acuity as is depicted in Figure 1 which depicts the fovea of the human eye.

This hierarchy plays a crucial role for the human attention mechanism as the wider representations contain low-resolution information of the environment that guides the more highly detailed foveal view towards the Subject-of-Interest (SOI) through the hierarchical processing of progressive views with ever increasing acuity.

An example of this occurs when something in the corner of the eye catches the person's attention and they move their eyes to achieve a more highly detailed view of the object. The eyes move to align and focus the fovea onto the SOI so that it can be seen with a high degree of visual acuity. While in the corner of the eye, the person may not be able to identify the object, or discern details, but once the object is within the foveal view of the eye, the object and details are easily identifiable due to the high degree of acuity of the fovea.

Disclosed herein is a method for an Intelligent Vision System that is able to function as good as, or better than, human vision and its attention mechanism. This Intelligent Vision System is based on a unique combination of Artificial Intelligence (AI) and commercially available cameras and servos.

To replace human vision with an Intelligent Vision System, an AI capability is desired to perceive the environment in a manner that is similar to humans. This must be done at resolutions with very high levels of detail of 60 to 120 pixels/degree or greater and it must also support visual hierarchical layering in varying degrees of visual acuity, just like human vision.

The improvement that this Intelligent Vision System commands over existing cameras and camera systems both with and without AI support is that it is based on the human vision model of hierarchical layering in varying degrees of visual acuity coupled with the human attention mechanism and a unique coupling of AI, cameras and servos.

For any other vision system to achieve this performance it would require an AI system with approximately 1,000 directional full-HD cameras to observe the 360˚ view of the environment. An AI system that can handle the input of 1,000 cameras in real-time is not practical for real-life applications due to extreme cost, high bandwidth requirements and implementation complexity.

Practical applications of the Intelligent Vision Systems would include autonomous driving, surveillance systems, law enforcement, security systems, improving the performance of identifying faces in a crowded environments such as airports, crowded cities, metro stations, events, and similar environments, reading auto and truck details and license plates from increased

distances, Automatic Plate Number Recognition (ANPR) for vehicle on-the-move systems, free movement expressway systems and more.

Summary

Accordingly, intelligent vision systems and methods for use in security/surveillance camera systems is desired. Intelligent vision systems and methods, for use in security/surveillance camera systems, that emulates the human retina and the attention mechanism of the human brain, replicating the central region of human vision ("Fovea™") and its surrounding hierarchical vision and attention system results in a significant performance increase over existing security/surveillance camera systems and any other camera system.

As disclosed and described herein, such intelligent vision systems and methods can be achieved through the innovative use of AI coupled with hardware, software, firmware, a combination or the like

To achieve the equivalent of human-level all-around high perception and vision acuity, deep learning AI is used to control one or more cameras with a field of view (FOV) that generates a set of Progressive Centrist Hierarchical Views (PCHV) of the visual environment. Each of the cameras can be equipped with a servo operated rapid pan/tilt mechanism that is controlled by a state-of-the-art attention mechanism to achieve the enhanced resolution, accurate movement and rapid focusing necessary to perceive the environment in high details.

The Intelligent Vision System state-of-the-art attention mechanism uses a controller mechanism with deep learning capabilities and algorithms. This achieves enhanced resolutions and improved attention driven focus, with resolutions that are greater than can be achieved by the human eye and with greater distance viewing capabilities and 360˚ surround vision.

Such intelligent vision systems and methods may consist of an all in one camera system that thinks and acts intelligently like a human and has the accuracy, efficiency, and speed for real-time applications with better than human sight.

Controlling a hierarchical view camera system is achieved by the disclosed attention mechanisms that include deep-learning, reinforcement learning, and human behavioral cloning imitation. The disclosed intelligent vision systems and methods provide a wider area of vision coverage, higher resolution, and further distance acuity than previously achieved by known systems. This includes

a 360° field of view coverage, with greater than human eye resolution (beyond 120 pixels/degree) providing up to 10 times (10x) the focal distance of other known solutions.

When monitoring a designated area, the disclosed intelligent vision systems and methods reduce the number of cameras required to cover the area. This is accomplished through the use of an advanced artificial intelligence system which drives the camera's rapid movement to the SOI utilizing an attention system.

Description of the Drawings

- Figure 1: Illustrates the fovea of the human eye and field of view.
- Figure 2: Illustrates a simulated human face used in the nine-ball environment.
- Figure 3: Provides an overview of the HRI attention game experiments.
- Figure 4(a): Provides a table summarizing the experimental results for camera and field of view configurations.
- Figure 4(b) – 4(d): Provides the learning graphs of the experiments provided in the table of Figure 4(a).
- Figure 5: Illustrates a block diagram with the main electro-mechanical components used in Sharp Vision.
- Figure 6: Provides a front profile of Sharp Vision prototype with the key electro-mechanical components highlighted.
- Figure 7: Provides a side profile of Sharp Vision prototype with the key electro-mechanical components highlighted.
- Figure 8: Illustrates a data flow diagram explaining the action-observation loop used in Sharp Vision.
- Figure 9: Illustrates a detailed data flow diagram of the action-observation loop used in Sharp Vision.
- Figure 10: Illustrates a more generic data flow diagram where multiple sensor streams can be processed for mirror control.

Detailed Description

Referring to Figures 1 through 10 an embodiment of the disclosed intelligent vision system and method is illustrated. The intelligent vision systems and methods herein are based on a foveal

view with an attention model that utilizes a hierarchical set of video image streams. Each video image stream includes an appropriate FOV to bring high resolution attention to the SOI.

Development of the disclosed intelligent vision systems and methods included seven Human-Robot Interaction (HRI) attention game experiments which were performed using one, two or three cameras with varying FOV. The subjects were nine balls with simulated human faces as shown in Figure 2. HRI is a challenging task that requires an AI Deep Learning Neural Network for detecting subtle visual cues such as an eye gaze, and the like.

The varying FOVs where achieved by changing the lenses between experiments. The balls had two circles drawn on them to represent eyes and a small cone was glued onto each ball below the eyes to represent a nose as is depicted in Figure 2.

Referring to Figure 3, the basic principles of the HRI attention game experiments are depicted. The camera(s) is situated on the bottom left of the HRI Attention game table and the nine simulated human faces (in this case annotated balls) are positioned at various locations within the FOV of the camera(s).

For each experiment there are a series of episodes in which a randomly selected face looks toward the camera(s) while the other faces will look in random directions. The task for the camera(s) AI is to locate the face that is looking at it. For each step, within the episode, the AI rotates the camera(s) by one degree in any direction. The results include two components: for each time step the AI receives a reward of +1 if it can locate the correct face or a punishment of -0.1 if it cannot. Each episode has a maximum of 50 steps.

During each experiment, the position of the human face balls was adjusted by either moving the faces towards the camera or away from the camera and measuring the time required for the AI-Deep Learning Neural Network to focus its attention on the ball with the simulated human face which is best facing the camera.

The goal of the experiments was to measure the time to achieve the attention focus for different numbers of cameras and varying FOVs.

Prior to starting the experiments, a baseline for a successful attention focus was established to set the maximum score, which is the best score that can be achieved.

The table of figure 4(a) summarizes the camera and FOV configurations for each experiment and the results of the experiments in the last two columns. Figures 4(b), 4(c) and 4(d) provide

graphical representation of the experimental results.

One camera is used for experiments 1, 2 and 3. For each of these experiments a single lens is used but with a progressively wider field of view ranging from narrow to wide. The outcome of the experiments for the HRI Game result with a field of view ranging from -0.73 to -0.60 are depicted graphically in Figure 4 (b). These results are with a single camera regardless of the FOV shows the lowest performance for achieving SOI attention focus.

Two cameras are used for experiments 4, 5 and 6. For each of these experiments one narrow angle lens is used on one camera and a wide-angle lens is used on the other camera; but for each experiment it is a different combination of lenses. The outcome of the experiments for the HRI Game results with a field of view ranging from -0.59 to -0.16 are depicted graphically in Figure 4 (c). These results are with a single camera regardless of the FOV and show an improved performance for achieving SOI attention focus as compared to experiments 1, 2 and 3. The use of two cameras has improved the convergence of the performance towards that of the baseline with experiment 6 achieving the best performance.

Three cameras are used for experiment 7. For this experiment, three lenses are used; one narrow angle lens is used on one camera, a midrange lens on another camera and a wide-angle lens is used on the third camera. The outcome of the experiments was an HRI Game result of 0.04 which is depicted graphically in Figure 4 (d). With a single camera regardless of the FOV improved performance is achieved for SOI attention focus as compared to experiments 1 through 6. The use of three cameras with appropriately selected lenses has further improved the convergence of the performance towards the baseline with experiment 7 achieving the best performance out of the seven experiments.

As illustrated in Figures 4(a) – (d), each task setting was trained through an AI agent with each of the foveal representations shown until convergence. As can also be seen in Figure 4, using the three set of parameters and the learning curves shown, the performance of all the experiments are summarized in the table of Figure 4(a). The values were computed by the average score over the last 100 trials. It can be seen that there are large variations in performance depending on the foveal representation used and it is a crucial hyper-parameter for the attention model.

In in Figures 4(a) – (d), it can be seen that the camera with FOV of 30 degree is a preferred representation for this setoff parameters as a single observation, while cameras with FOV of 10 degrees and 60 degrees fail. This indicates that with a single observation, the FOV can greatly

affect performance. Experiments 4 and 5, depicted in Figure 4 (c) and the table of figure 4(a), achieve low performance as the combination of two observations have large variations in FOV and the wider camera is unable to guide the narrower camera towards the SOI. On the other hand, experiment 6 has better performance as both cameras are able to work together efficiently by forming a hierarchy that improves ability to zero in of the SOI. Finally, from Figure 4(c) it can be seen that increasing the number of cameras increases performance as the cameras capture information from the environment at an increasing number of hierarchical views with improved resolutions.

Similar performance is achieved with the image classification task as shown in Figure 4(d) with experiment number 7 having the best performance in both cases. However, there were slight variations in comparison to the HRI task such as experiments with three cameras being better than experiments with one or two cameras. From these experiments it can be seen that the AI Deep Learning neural network needs to focus on global images features and operation is improved with a wider FOV in general.

To train the attention model, Proximal Policy Optimization (PPO) was used. PPO is an efficient policy search algorithm capable of handling high-dimensional observations. The policy function is represented by a Dynamic Neural Network (DNN) as shown in Figures 4(a) – (d). The input to the network includes both the environment state and observations.

The state representation is given by s_t R^{2nc} with n_c as the number of cameras and the state for each camera as its current orientation along two axes of rotation. Each camera also provides an observation in the form of an RGB image $o^{(ci)}$ $\mathrm{R}^{64\times64\times3}$.

There are two outputs generated by the AI Deep Learning neural network which include the policy and a scalar state value estimate. The policy output is a “softmax” probability distribution $\pi_\theta(a_t\ s_t)$ R^9 where the output actions drive the motion of the cameras along one of the eight cardinal directions.

Figures 5 through 10 illustrate one preferred embodiment of the active vision system for achieving sharp vision described herein. This preferred embodiment in Figures 5-7 includes a head (10) to which are mounted cameras (12), (13) and (14).

A mirror (11) with a pan/tilt mechanism is included above the camera lenses to control the direction of the cameras view.

Mirror 11 includes a fast real-time actuator (22) with a movement frequency of 10 actions/second. Fast real-time actuator (22) is controlled by processor (23). This provides for rapid movement and rotation of each camera view (12), (13), (14) to desired specific pan/tilt angles and locations.

While three cameras are illustrated in this embodiment it will be appreciated by those skilled in the art that the Intelligent Vision Systems and Methods can include any number of cameras, camera combinations, camera proximity, camera network (IP or otherwise) or integrated cameras (i.e. cameras with multiple sensors) which may be used to realize the desired and improved Intelligent Vision Systems and Methods.

The images shown in Figure 6 and Figure 7 are photographs of the actual prototype.

The active vision system for achieving sharp vision utilizes a controller (Figure 6 -23) for performing the real-time control of the servos (where is figures 6 22) and cameras. The controller is hardware with firmware that is embedded in the hardware. The controller utilizes the output from the artificial intelligence-deep learning neural network algorithm to manage the positioning of the pan/tilt cameras by issuing commands to the servos.

Servo driven mechanical movement is achieved using gimbals, gearing, pivots mechanical movement devices. The real-time rotation of the end-effector pan/tilt mechanisms is achieved with fast servos.

The interaction between cameras (81, 82, …n) image processing, image streams, neural networks, controllers, pan/tilt functionality and the repeating of one or more of these functionalities is achieved by the controller firmware. In each process module or functionality, it is necessary to execute a process flow which includes one or more of the following steps which are shown in the flowchart depicted in Figure 8:

1. capturing images from one or more cameras (83);
2. processing the images to generate the required streams (84);
3. feeding forward or otherwise making the images available to the neural network (85);
4. reading the output from the neural network to control the pan/tilt position or other positioning of one or more cameras for the next time step (86); or

Start again from step 1 to train the system and further the deep learning process. All these steps are happening simultaneously and rapidly.

The active vision system for achieving sharp vision leverages Open Source software for image capture and for preprocessing the images.

1. (Open Source Computer Vision Library) OpenCV™ for image capture and processing

   OpenCV provides a common infrastructure for computer vision applications and machine perception. This software library can be used to detect and recognize faces, identify objects, classify human actions in videos, track camera movements, track moving objects, extract 3D models of objects, produce 3D point clouds from stereo cameras, stitch images together to produce a high resolution image of an entire scene, find similar images from an image database, remove red eyes from images taken using flash, follow eye movements, recognize scenery and establish markers to overlay it with augmented reality, etc.

2. Tensorflow™ is used for control and to construct and train the neural network to produce the desired behavior thus achieving the desired classification outputs.

   TensorFlow is an end-to-end open source platform for machine learning supported numerical computations that makes machine learning faster and easier. TensorFlow provides an abstraction that includes machine learning and deep learning (i.e. neural networking) models and algorithms and manages the various options for directing the output of one function to the input of another.

3. Robot Operating System (ROS™) to provide the servo control of the pan/tilt mechanism in real-time

   ROS is a flexible robot software framework. It is a collection of tools, libraries, and conventions that aim to simplify the task of creating complex and robust robot behavior across a wide variety of robotic platforms.

The active vision system for achieving sharp vision disclosed includes software with computer vision code to read the images from the cameras and process them in real-time.

The image shown in Figure 9 depicts all of these elements coming together to realize the intelligent vision system.

The input images from the cameras are divided into multiple streams using a 90˚ (for example) FOV camera which can capture the center part of the image that represent 30˚ (for example) FOV

then resize both images to a specific resolution to feed a deep controller neural network. Such configuration will reduce the computational demand significantly without losing fine details in the 30˚ FOV.

As illustrated in Figure 10, each camera sensor may provide a signal that generates a stream of data as detected by the sensor. Such sensors can be one or more of the following:

- Zoom camera sensor (14) develops the zoom camera stream (24);
- Wide camera sensor (12) develops wide camera stream (26);
- Surround camera sensor (13) develops surround camera stream (28).

The zoom camera (24), wide camera (26) and surround camera (26) streams may be fed into software level streams at level (0) to level (4). Such feeds may include:

- Zoom camera stream (24) is fed into software level streams 0 and 1 (28), (30)
- Wide camera stream (26) is fed into software level streams 2 and 3 (32), (34)
- Surround camera stream (28) is fed into software level stream 4 (36).

Software level stream 0 to software level stream 4 are processed by deep learning layers (38, 40, 42, 44 and 46). These deep learning layers (38, 40, 42, 44 and 46) are concatenated (48) and fed to long short-term memory (LSTM) modules T1, T2, T3… Tn (50, 52, 54, 56). The outputs (58, 60, 62, 64) of LSTM modules T1 to Tn (50, 52, 54, 56) can be classified and placed into memory (66, 68, 70 and 72) for future use.

LSTM outputs (66, 68, 70, 72) provide feedback through feedback loop (74) to adjust the pan and tilt mechanism controlled by computational processor (23). A configuration of lenses (18), (19) and (20) with multiple progressive centrist FOV parameters are used to provide the desired focal lengths and properties for the Intelligent Vision System and Methods. Such FOV parameters include but are not limited to, 360˚, 120˚, 10˚ which provides for human level resolution of approximately 120 pixels/degree.

Other embodiments may include, but are not limited to, FOV parameters of 360˚,120˚,60˚,10˚,2˚ which provides beyond human level resolution of more than 120 pixels/degree.

Reference throughout this specification to “the embodiment,” “this embodiment,” “the previous embodiment,” “one embodiment,” “an embodiment,” “a preferred embodiment” “another preferred embodiment” or similar language means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one

embodiment of the present invention. Thus, appearances of the phrases “in the embodiment, “in this embodiment,” “in the previous embodiment, in one embodiment, in an embodiment,” “in a preferred embodiment,” “in another preferred embodiment,” and similar language throughout this specification may, but do not necessarily, all refer to the same embodiment.

The described features, advantages, and characteristics of the invention may be combined in any suitable manner in one or more embodiments. One skilled in the relevant art will recognize that the invention may be practiced without one or more of the specific features or advantages of a particular embodiment. In other instances, additional features and advantages may be recognized in certain embodiments that may not be present in all embodiments of the invention. While the present invention has been described in connection with certain exemplary or specific embodiments, it is to be understood that the invention is not limited to the disclosed embodiments, but, on the contrary, is intended to cover various modifications, alternatives, modifications and equivalent arrangement as will be apparent to those skilled in the art. Any such changes, modifications, alternative, equivalents and the like may be made without departing from the spirit and scope of the disclosure.

CLAIMS

1. An intelligent vision system comprising:

   a. a camera positioned for creating an image data stream output for the camera and capturing images from the camera;

   wherein the camera is a fixed-angle camera operably aligned with a mirror mounted on a pan and tilt real-time actuator mechanism operable to obtain a desired angle and location view on the mirror which are captured as images by the camera;

   b. a processor for processing the image data stream output of the camera to generate one or more image data streams for the camera;

   c. a neural network for receiving and further processing the image data stream output from the camera;

   d. a reader for reading the output from the neural network to control the position of the camera; and

   e. a feedback network for further processing the image data stream output of the camera to ultimately focus in on the Subject of Interest (SOI).

2. The intelligent vision system of claim 1, further including an artificial intelligence neural network and a second (or more) camera, wherein the artificial neural network manages a

hierarchy of different view angles generated by the first camera and the second (or more) camera.

3. The intelligent vision system of claim 2, wherein the first and second (or more) cameras function as a single camera system.

4. The intelligent vision system of claim 1, wherein the artificial intelligence neural network includes a deep learning module controlling rapid movement of the cameras to find, detect or track one or more SOI targets.

5. The intelligent vision system of claim 1, wherein the camera produces a normal resolution image data stream and the processor provides a high-resolution image data stream of one or more targets, wherein the normal resolution image data stream and the high-resolution image data stream are displayed together.

6. A method of providing intelligent vision images including using the intelligent vision system of claim 1, said method including the steps of :

   a. positioning at least one camera for creating an image data stream output for each camera;

   b. capturing image data stream output from the at least one camera;

   c. processing the image data stream output from the at least one camera to generate one or more data streams;

   d. further processing the image data stream output from the at least one camera through a neural network;

   e. reading the output from the neural network to control the position of the at least one camera; and

   f. further processing the image data stream of the at least one camera through a feedback network.

7. The method of claim 6, further positioning a second (or more) camera for creating a second image data stream output, and an artificial neural network managing a hierarchy of different view angles generated by the first camera and the second (or more) camera.

8. The method of claim 7, wherein the first and second (or more) cameras function as a single camera system.

9. The method of claim 6, wherein further processing the image data stream of the at least one camera through a feedback network includes processing the image data stream of the at least one camera through a deep learning module controlling rapid movement of the camera to find, detect or track one or more SOI targets.

10.

The method of claim 6, wherein processing the image data stream output of the at least one camera to generate one or more data streams produces a normal resolution image data stream and further processing the image data stream output from the at least one camera through a neural network provides a high-resolution image data stream of one or more SOI targets, wherein the normal resolution image data stream and the high-resolution image data stream are displayed together.

الوصف

نظام رؤية فعال لتحقيق رؤية حادة

<u>الخلفية التقنية:</u>

تُعرَّف حدة البصر بأنها الحدة أو الوضوح في الرؤية. يدرك البشر بصريًا بيئتهم بمستويات عالية جدًا من التفاصيل بدقة عالية جدًا (60 إلى 120 بكسل/درجة) في بيئة 360 درجة. تسمح حدة البصر هذه للبشر بأداء مهام مثل القيادة والمشاركة في التفاعلات الاجتماعية التي تتطلب الانتباه إلى الإشارات المرئية الدقيقة مثل قصد السائر على قدميه أو تعبيرات الوجه.

تعد حدة البصر هي وظيفة النقرة التي تكون جزء من العين التي لها القدرة على تحقيق أقصى حدة. في مركز النقرة، تكون حدة البصر في ذروتها. بالابتعاد عن النقرة، تتضاءل حدة البصر تدريجيًا، وبالتالي إنشاء نموذج تمثيل مرئي عبارة عن طبقات هرمية بدرجات متفاوتة من الحدة كما هو موضح في الشكل 1 الذي يصور نقرة العين البشرية.

يلعب هذا التسلسل الهرمي دورًا حاسمًا في آلية الانتباه البشري حيث أن التمثيلات الأوسع تحتوي على معلومات منخفضة الدقة عن البيئة التي توجه رؤية نقيرية أكثر تفصيلًا نحو الهدف موضوع الاهتمام (SOI) من خلال المعالجة الهرمية للرؤى التقدمية مع حدة متزايدة.

يحدث مثال على ذلك عندما يلفت شيء ما في زاوية العين انتباه الشخص ويحرّك عينيه لتحقيق رؤية أكثر تفصيلًا للهدف. تتحرك العيون لمحاذاة وتركيز النقرة على SOI بحيث يمكن رؤيتها بدرجة عالية من حدة البصر. أثناء وجوده في زاوية العين، قد لا يكون الشخص قادرًا على تحديد الهدف، أو تمييز التفاصيل، ولكن بمجرد أن يكون الهدف في نطاق الرؤية النقية للعين، يمكن التعرف على الهدف والتفاصيل بسهولة بسبب درجة الحدة العالية للنقرة.

تم الكشف هنا عن طريقة لنظام رؤية ذكي قادر على العمل بجودة، أو أفضل من، الرؤية البشرية وآلية الانتباه الخاصة بها. يعتمد نظام الرؤية الذكي هذا على مزيج فريد من الذكاء الاصطناعي (AI) والكاميرات والمحركات المؤازرة المتاحة تجاريًا.

لاستبدال الرؤية البشرية بنظام رؤية ذكي، فإن قدرة الذكاء الاصطناعي تكون مطلوبة لإدراك البيئة بطريقة تشبه البشر. يجب أن يتم ذلك بدقة مع مستويات عالية جدًا من التفاصيل من 60 إلى 120 بكسل/درجة أو أكثر، ويجب أيضًا أن يدعم الطبقات الهرمية المرئية بدرجات متفاوتة من حدة البصر، تمامًا مثل الرؤية البشرية.

يعد التحسين الذي يأمر به نظام الرؤية الذكي هذا على الكاميرات وأنظمة الكاميرا الحالية مع وبدون دعم AI هو أنه يعتمد على نموذج الرؤية البشرية للطبقات الهرمية بدرجات متفاوتة من حدة البصر إلى جانب آلية الانتباه البشري والاقتران الفريد بين AI، الكاميرات والمحركات المؤازرة.

لكي يحقق أي نظام رؤية آخر هذا الأداء، سيتطلب نظام ذكاء اصطناعي به ما يقرب من 1000 كاميرا عالية الدقة اتجاهية لمراقبة رؤية البيئة بزاوية 360 درجة. إن نظام AI الذي يمكنه التعامل مع دخل لـ 1000 كاميرا في الوقت الفعلي ليس عمليًا للتطبيقات الواقعية بسبب التكلفة الباهظة ومتطلبات النطاق الترددي العالي وتعقيد التنفيذ.

تشمل التطبيقات العملية لأنظمة الرؤية الذكية القيادة الذاتية، وأنظمة المراقبة وتنفيذ القانون، وأنظمة الأمن، وتحسين أداء التعرف على الوجوه في البيئات المزدحمة مثل المطارات، والمدن المزدحمة، ومحطات المترو، والأحداث، والبيئات المماثلة، قراءة تفاصيل السيارات والشاحنات ولوحات الترخيص من مسافات متزايدة، والتعرف التلقائي على رقم اللوحة (ANPR) لأنظمة المركبات أثناء التنقل، وأنظمة الطرق السريعة ذات الحركة الحرة والمزيد.

<u>الكشف عن الاختراع:</u>

وبناءً على ذلك، فإن أنظمة الرؤية الذكية وطرق استخدامها في أنظمة كاميرات المراقبة/الأمن تكون مرغوبة. تؤدي أنظمة وطرق الرؤية الذكية، لاستخدامها في أنظمة كاميرات المراقبة/الأمن، والتي تحاكي شبكية العين البشرية وآلية الانتباه للدماغ البشري، تكرار المنطقة المركزية للرؤية البشرية ("Fovea™") ونظام الرؤية الهرمية المحيطة بها ونظام الانتباه إلى زيادة كبيرة في الأداء مقارنة بأنظمة كاميرات الحماية/المراقبة الحالية وأي نظام كاميرا آخر.

كما تم الكشف عنه ووصفه هنا، يمكن تحقيق أنظمة وطرق الرؤية الذكية من خلال الاستخدام المبتكر لـ AI إلى جانب الأجهزة، البرامج، البرامج الثابتة، توليفة أو ما شابه ذلك.

لتحقيق ما يعادل الإدراك العالي الشامل على المستوى البشري وحدة الرؤية، يتم استخدام التعلم العميق AI للتحكم في واحدة أو أكثر من الكاميرات بمجال الرؤية (FOV) الذي يولد مجموعة من الرؤى الهرمية المركزية التقدمية (PCHV) من البيئة المرئية. يمكن تجهيز كل من الكاميرات بآلية دوران/إمالة سريعة تعمل بمؤازرة يتم التحكم فيها بواسطة آلية انتباه متطورة لتحقيق الدقة المحسنة، والحركة الدقيقة، والتركيز السريع الضروري لإدراك البيئة بتفاصيل عالية.

تستخدم آلية الانتباه المتطورة لنظام الرؤية الذكية آلية تحكم ذات قدرات وخوارزميات التعلم العميق. يحقق هذا دقة محسّنة وتركيزًا محسّنًا مدفوعًا بالاهتمام، مع دقة أكبر مما يمكن تحقيقه بالعين البشرية وبإمكانيات رؤية أكبر عن بُعد ورؤية محيطية بزاوية 360 درجة.

قد تتكون أنظمة وأساليب الرؤية الذكية هذه من نظام كاميرا شامل يفكر ويعمل بذكاء مثل الإنسان ولديه الدقة والكفاءة والسرعة للتطبيقات في الوقت الفعلي مع رؤية أفضل من الإنسان.

يتحقق التحكم في نظام كاميرا الرؤية الهرمية من خلال آليات الانتباه التي تم الكشف عنها والتي تشمل التعلم العميق، والتعلم المعزز، وتقليد الاستنساخ السلوكي البشري. توفر أنظمة وأساليب الرؤية الذكية التي تم الكشف عنها مساحة أوسع لتغطية الرؤية، ودقة أعلى، ودقّة مسافة أكبر مما حققته الأنظمة المعروفة سابقًا. يتضمن ذلك تغطية مجال الرؤية بزاوية 360 درجة، مع دقة أعلى من دقة العين البشرية (تتجاوز 120 بكسل/درجة) توفر ما يصل إلى 10 أضعاف (10 مرات) المسافة البؤرية من الحلول المعروفة الأخرى.

عند مراقبة مساحة معينة، تعمل أنظمة وأساليب الرؤية الذكية التي تم الكشف عنها على تقليل عدد الكاميرات المطلوبة لتغطية المساحة. يتم تحقيق ذلك من خلال استخدام نظام ذكاء اصطناعي متقدم يدفع حركة الكاميرا السريعة إلى SOI باستخدام نظام الانتباه.

<u>وصف الرسومات</u>

- الشكل 1 يوضح نقرة العين البشرية ومجال الرؤية.
- الشكل 2 يوضح محاكاة لوجه بشري مستخدم في بيئة من تسع كرات.
- الشكل 3 يوضح لمحة عامة عن تجارب لعبة الانتباه HRI.
- الشكل 4 (أ) عبارة عن جدول يلخص النتائج التجريبية لتهيئات الكاميرا ومجال الرؤية.
- الأشكال 4 (ب) - 4 (ج) توضح الرسوم البيانية للتجارب الموضحة في الجدول للشكل 4 (أ).
- الشكل 5 يوضح رسماً تخطيطياً مع المكونات الكهروميكانيكية الرئيسية المستخدمة في رؤية حادة.
- الشكل 6 يوضح صورة أمامية لنموذج أولي لرؤية حادة مع إبراز المكونات الكهروميكانيكية الرئيسية.
- الشكل 7 يوضح صورة جانبية لنموذج أولي لرؤية حادة مع إبراز المكونات الكهروميكانيكية الرئيسية.
- الشكل 8 يوضح رسم بياني لتدفق البيانات يشرح حلقة مراقبة العمل المستخدمة في رؤية حادة.
- الشكل 9 يوضح مخططاً تفصيلياً لتدفق البيانات لحلقة مراقبة العمل المستخدمة في رؤية حادة.
- الشكل 10 يوضح مخططاً أكثر عمومية لتدفق البيانات حيث يمكن معالجة تدفقات أجهزة الاستشعار المتعددة للتحكم في المرآة.

<u>الوصف المفصل للاختراع:</u>

بالإشارة إلى الأشكال من 1 إلى 10، تم توضيح نموذج لنظام وطريقة الرؤية الذكية التي تم الكشف عنها. تعتمد أنظمة وطرق الرؤية الذكية هنا على رؤية نقَيْرِيَّة بنموذج انتباه يستخدم مجموعة هرمية من تدفقات صور الفيديو. يتضمن كل تدفق لصور الفيديو FOV مناسب لجذب الانتباه عالي الدقة إلى SOI.

يشتمل تطوير أنظمة وطرق الرؤية الذكية التي تم الكشف عنها على سبع تجارب على ألعاب الانتباه للتفاعل بين الإنسان والروبوت (HRI) والتي تم إجراؤها باستخدام كاميرا واحدة أو اثنتين أو ثلاث كاميرات مع FOV مختلف. كانت الأهداف تسع كرات تحاكي وجود بشرية كما هو موضح في الشكل 2. تعد HRI مهمة صعبة تتطلب الشبكة العصبية للتعلم العميق AI لاكتشاف الإشارات المرئية الدقيقة مثل نظرة العين، وما شابه ذلك.

حيث يتم تحقيق FOVs المتغيرة عن طريق تغيير العدسات بين التجارب. تم رسم دائرتين على الكرات لتمثيل العينين وتم لصق مخروط صغير على كل كرة أسفل العينين لتمثيل الأنف كما هو موضح في الشكل 2.

بالإشارة إلى الشكل 3، تم وصف المبادئ الأساسية لتجارب لعبة انتباه HRI. توجد الكاميرا (الكاميرات) في الجزء السفلي الأيسر من جدول لعبة انتباه HRI ويتم وضع الوجوه البشرية التسعة المحاكية (في هذه الحالة الكرات المعينة) في مواقع مختلفة داخل مجال الرؤية للكاميرا (الكاميرات).

لكل تجربة، هناك سلسلة من النوبات التي يتجه فيها الوجه المختار عشوائيًا نحو الكاميرا (الكاميرات) بينما تنظر الوجوه الأخرى في اتجاهات عشوائية. تكون مهمة الكاميرا (الكاميرات) AI هي تحديد الوجه الذي ينظر إليها. لكل خطوة، داخل النوبة، يقوم AI بتدوير الكاميرا (الكاميرات) بدرجة واحدة في أي اتجاه. تتضمن النتائج عنصرين: لكل خطوة زمنية، يتلقى AI مكافأة +1 إذا كان بإمكانه تحديد الوجه الصحيح أو عقوبة -0.1 إذا لم يتمكن من ذلك. تحتوي كل نوبة على 50 خطوة كحد أقصى.

أثناء كل تجربة، تم تعديل موضع كرات الوجه البشري إما عن طريق تحريك الوجوه نحو الكاميرا أو بعيدًا عن الكاميرا وقياس الوقت اللازم لشبكة AI للتعلم العميق لتركيز انتباهها على الكرة مع الوجه البشري المحاكي الذي هو أفضل مواجهة للكاميرا.

كان الهدف من التجارب هو قياس الوقت اللازم لتحقيق تركيز الانتباه لأعداد مختلفة من الكاميرات وFOVs المتباينة.

قبل بدء التجارب، تم إنشاء خط أساس لتركيز الانتباه الناجح لتعيين الحد الأقصى للدرجة، وهي أفضل درجة يمكن تحقيقها.

يلخص جدول الشكل 4 (أ) الكاميرا وتهيئات FOV لكل تجربة ونتائج التجارب في العمودين الأخيرين. توفر الأشكال 4 (ب) و4 (ج) و4 (د) تمثيلًا رسوميًا للنتائج التجريبية.

يتم استخدام كاميرا واحدة للتجارب 1 و2 و3. لكل من هذه التجارب، يتم استخدام عدسة واحدة ولكن مع مجال رؤية أوسع بشكل تدريجي يتراوح من ضيق إلى واسع. تم توضيح نتائج التجارب الخاصة بنتيجة لعبة HRI مع مجال رؤية يتراوح من -0.73 إلى -0.60 بيانيًا في الشكل 4 (ب). تكون هذه النتائج مع كاميرا واحدة بغض النظر عن أن FOV يظهر أقل أداء لتحقيق بؤرة انتباه SOI.

يتم استخدام كاميرتين للتجارب 4 و5 و6. لكل من هذه التجارب، يتم استخدام عدسة واحدة ضيقة الزاوية في كاميرا واحدة ويتم استخدام عدسة واسعة الزاوية في الكاميرا الأخرى؛ ولكن لكل تجربة تكون عبارة عن توليفة مختلفة من العدسات. تم توضيح نتائج التجارب الخاصة بنتائج لعبة HRI مع مجال رؤية يتراوح من -0.59 إلى -0.16 بيانيًا في الشكل 4 (ج). تكون هذه النتائج مع كاميرا واحدة بغض النظر عن FOV وتظهر أداء محسنًا لتحقيق بؤرة انتباه SOI مقارنة بالتجارب 1 و2 و3. أدى استخدام كاميرتين إلى تحسين تقارب الأداء مع أداء القيمة الأساسية مع التجربة 6 لتحقيق أفضل أداء.

تم استخدام ثلاث كاميرات للتجربة 7. في هذه التجربة، تم استخدام ثلاث عدسات؛ يتم استخدام إحدى العدسات ذات الزاوية الضيقة على إحدى الكاميرات، وتستخدم عدسة متوسطة المدى على كاميرا أخرى وعدسة واسعة الزاوية في الكاميرا الثالثة. كانت حصيلة التجارب نتيجة لعبة HRI بقيمة 0.04 والتي تم تصويرها بيانيًا في الشكل 4 (د). باستخدام كاميرا واحدة بغض النظر عن FOV، يتم تحقيق الأداء المحسن لبؤرة انتباه SOI مقارنة بالتجارب من 1 إلى 6. وقد أدى استخدام ثلاث كاميرات مع عدسات مختارة بشكل مناسب إلى تحسين تقارب الأداء نحو القيمة الأساسية مع تحقيق التجربة 7 لتحقيق أفضل أداء من بين التجارب السبعة.

كما هو موضح في الأشكال 4 (أ) - (د)، تم التدريب كل إعداد مهمة بواسطة عامل AI مع كل تمثيلات نقيرية معروضة حتى التقارب. كما يمكن رؤيته أيضًا في الشكل 4، باستخدام المجموعة الثلاث من المتغيرات ومنحنيات التعلم الموضحة، يتم تلخيص أداء جميع التجارب في الجدول في الشكل 4 (أ). تم حساب القيم من خلال متوسط الدرجات خلال آخر 100 تجربة. يمكن ملاحظة أن هناك اختلافات كبيرة في الأداء اعتمادًا على التمثيل النقري المستخدم وهو متغير مفرط حاسم لنموذج الانتباه.

في الأشكال 4 (أ) - (د)، يمكن ملاحظة أن الكاميرا ذات FOV 30 درجة هي تمثيل مفضل لمتغيرات الإعداد هذه كملاحظة واحدة، بينما تفشل الكاميرات ذات مجال الرؤية 10 درجات و60 درجة. يشير هذا إلى أنه من خلال ملاحظة واحدة، يمكن أن يؤثر FOV بشكل كبير على الأداء. تحقق التجارب 4 و5، المصورة في الشكل 4 (ج) وجدول الشكل 4 (أ)، أداءً منخفضًا لأن الجمع بين ملاحظتين لهما اختلافات كبيرة في FOV والكاميرا الأوسع غير قادرة على توجيه الكاميرا الأضيق نحو SOI. من ناحية أخرى، تتمتع التجربة 6 بأداء أفضل حيث أن كلتا الكاميرتين قادرة على العمل معًا بكفاءة من خلال تشكيل تسلسل هرمي يحسن القدرة على الصفر في SOI. أخيرًا، من الشكل 4 (ج)، يمكن ملاحظة أن زيادة عدد الكاميرات تزيد من الأداء حيث تلتقط الكاميرات المعلومات من البيئة بعدد متزايد من العروض الهرمية مع دقة محسنة.

يتم تحقيق أداء مماثل مع مهمة تصنيف الصور كما هو موضح في الشكل 4 (د) مع التجربة رقم 7 التي تتمتع بأفضل أداء في كلتا الحالتين. ومع ذلك، كانت هناك اختلافات طفيفة بالمقارنة مع مهمة HRI مثل التجارب مع ثلاث كاميرات أفضل من التجارب باستخدام كاميرا واحدة أو اثنتين. من هذه التجارب، يمكن ملاحظة أن الشبكة العصبية للتعلم العميق بـ AI تحتاج إلى التركيز على ميزات الصور العالمية وتم تحسين التشغيل باستخدام FOV أوسع بشكل عام.

لتدريب نموذج الانتباه، تم استخدام تحسين سياسة القرب (PPO). تعد PPO هي خوارزمية بحث سياسة فعالة قادرة على التعامل مع الملاحظات عالية الأبعاد. يتم تمثيل وظيفة السياسة بواسطة الشبكة العصبية الديناميكية (DNN) كما هو موضح في الأشكال 4 (أ) - (د). تتضمن المدخلات إلى الشبكة كلًا من حالة البيئة والملاحظات.

يتم تقديم تمثيل الحالة بواسطة $\mathrm{R}^{2nc_{s_t}}$ مع n_c على أنه عدد الكاميرات والحالة لكل كاميرا كتوجهها الحالي على محورين للدوران. توفر كل كاميرا أيضًا ملاحظة في شكل صورة RGB من نوع $\mathrm{R}^{64\times64\times3(ci)}{}_o$.

هناك نوعان من النواتج التي تم إنشاؤها بواسطة الشبكة العصبية للتعلم العميق بـ AI والتي تتضمن السياسة وتقدير قيمة الحالة العددية. يعد ناتج السياسة هو توزيع احتمالي "softmax" - $(a_t\ s_t)\pi\theta$ R^9 حيث تدفع إجراءات المخرجات حركة الكاميرات على طول أحد الاتجاهات الأساسية الثمانية.

توضح الأشكال من 5 إلى 10 أحد النماذج المفضلة لنظام الرؤية الفعال لتحقيق رؤية حادة موصوفة هنا. يشتمل هذا النموذج المفضل في الأشكال 5-7 على رأس (10) مثبت عليه كاميرات (12)، (13) و(14).

يتم تضمين مرآة (11) بآلية دوران/إمالة فوق عدسات الكاميرا للتحكم في اتجاه عرض الكاميرا.

تتضمن المرآة 11 مشغلًا سريعًا في الوقت الفعلي (22) بتردد حركة يبلغ 10 إجراءات/ثانية. يتم التحكم في مشغل الوقت الحقيقي السريع (22) بواسطة المعالج (23). يوفر هذا الحركة السريعة والدوران لكل رؤية كاميرا (12)، (13)، (14) إلى زوايا ومواقع الدوران/الإمالة المحددة المطلوبة.

بينما يتم توضيح ثلاث كاميرات في هذا النموذج، سيكون موضع إدراك أولئك المهرة في المجال أن أنظمة وطرق الرؤية الذكية يمكن أن تتضمن أي عدد من الكاميرات أو مجموعات الكاميرا أو قرب الكاميرا أو شبكة الكاميرا (IP أو غير ذلك) أو الكاميرات المدمجة (أي الكاميرات ذات المستشعرات المتعددة) والتي يمكن استخدامها لتحقيق أنظمة وطرق الرؤية الذكية المرغوبة والمحسنة.

تعد الصور الموضحة في الشكل 6 والشكل 7 هي صور فوتوغرافية للنموذج الأولي الفعلي.

يستخدم نظام الرؤية الفعال لتحقيق رؤية حادة وحدة تحكم (الشكل 6 - 23) لأداء التحكم في الوقت الفعلي في المحركات المؤازرة (حيث الأشكال 6 - 22) والكاميرات. تكون وحدة التحكم عبارة عن جهاز به برنامج ثابت مضمن في الجهاز. تستخدم وحدة التحكم الخرج من خوارزمية الشبكة العصبية للتعلم العميق للذكاء الاصطناعي لإدارة وضع كاميرات الدوران/الإمالة عن طريق إصدار أوامر إلى المحركات المؤازرة.

يتم تحقيق الحركة الميكانيكية المؤازرة باستخدام وسائل الحركة الميكانيكية المحورية، والتروس، والمحاور. يتم تحقيق الدوران في الوقت الفعلي لآليات دوران/إمالة المستجيب النهائي باستخدام المحركات المؤازرة السريعة.

يتم تحقيق التفاعل بين معالجة صور الكاميرات (81، 82،... n)، وتدفق الصور والشبكات العصبية، ووحدات التحكم، ووظيفة الدوران/الإمالة وتكرار واحدة أو أكثر من هذه الوظائف بواسطة البرنامج الثابت لوحدة التحكم. في كل وحدة نمطية أو مجموعات وظيفية، من الضروري تنفيذ تدفق العملية الذي يتضمن خطوة واحدة أو أكثر من الخطوات التالية التي تظهر في المخطط الانسيابي الموضح في الشكل 8:

1. التقاط الصور من كاميرا واحدة أو أكثر (83)؛

2. تجهيز الصور لتوليد التدفقات المطلوبة (84).

3. التغذية إلى الأمام أو جعل الصور متاحة للشبكة العصبية (85)؛

4. قراءة الخرج من الشبكة العصبية للتحكم في موضع الدوران/الإمالة أو وضع آخر لكاميرا واحدة أو أكثر للخطوة الزمنية التالية (86)؛ أو

ابدأ مرة أخرى من الخطوة 1 لتدريب النظام وتعزيز عملية التعلم العميق. تحدث كل هذه الخطوات في وقت واحد وبسرعة.

يعمل نظام الرؤية الفعال لتحقيق رؤية حادة على تعزيز برامج المصدر المفتوح لالتقاط الصور والمعالجة المسبقة للصور.

1. (مكتبة رؤية الحاسوب مفتوحة المصدر) ™OpenCV لالتقاط الصور ومعالجتها

يوفر OpenCV بنية أساسية مشتركة لتطبيقات رؤية الحاسوب وإدراك الآلة. يمكن استخدام مكتبة البرامج هذه لاكتشاف الوجوه والتعرف عليها، وتحديد الأهداف، وتصنيف الأفعال البشرية في مقاطع الفيديو، وتتبع حركات الكاميرا، وتتبع الأجسام المتحركة، واستخراج نماذج ثلاثية الأبعاد للأهداف، وإنتاج سحب نقطية ثلاثية الأبعاد من كاميرات الاستريو، وتجميع الصور معًا لإنتاج صورة عالية صورة ذات دقة وضوح لمشهد كامل، والعثور على صور مماثلة من قاعدة بيانات الصور، وإزالة العيون الحمراء من الصور الملتقطة باستخدام الفلاش، واتباع حركات العين، والتعرف على المشهد وإنشاء علامات لتغطيته بالواقع المعزز، وما إلى ذلك.

2. يستخدم ™Tensorflow للتحكم ولإنشاء وتدريب الشبكة العصبية لإنتاج السلوك المطلوب وبالتالي تحقيق مخرجات التصنيف المرغوبة.

يعد TensorFlow عبارة عن نظام أساسي مفتوح المصدر شامل للحسابات الرقمية المدعومة بالتعلم الآلي والتي تجعل التعلم الآلي أسرع وأسهل. يوفر TensorFlow تجريدًا يتضمن نماذج وخوارزميات التعلم الآلي والتعلم العميق (أي الشبكات العصبية) ويدير الخيارات المختلفة لتوجيه ناتج إحدى الوظائف لدخل آخر.

3. نظام تشغيل الروبوت (™ROS) لتوفير التحكم المؤازر لآلية الدوران/الإمالة في الوقت الحقيقي

يعد ROS هو إطار عمل برنامج روبوت مرن. ويكون مجموعة من الأدوات، والمكتبات، والاتفاقيات التي تهدف إلى تبسيط مهمة إنشاء سلوك روبوت معقد وقوي عبر مجموعة واسعة من المنصات الآلية.

يشتمل نظام الرؤية الفعال لتحقيق الرؤية الحادة الذي يتم الكشف عنه على برنامج برمز رؤية الحاسوب لقراءة الصور من الكاميرات ومعالجتها في الوقت الفعلي.

توضح الصورة الموضحة في الشكل 9 كل هذه العناصر مجتمعة لتحقيق نظام الرؤية الذكي.

تنقسم الصور المدخلة من الكاميرات إلى تدفقات متعددة باستخدام 90 درجة (على سبيل المثال) كاميرا FOV التي يمكنها التقاط الجزء المركزي من الصورة التي تمثل 30 درجة (على سبيل المثال) FOV ثم تغيير حجم كلتا الصورتين إلى دقة محددة لتغذية شبكة عصبية لوحدة تحكم عميقة. ستؤدي هذه التهيئة إلى تقليل متطلبات الحسابات بشكل كبير دون فقدان التفاصيل الدقيقة في 30 FOV درجة.

كما هو موضح في الشكل 10، قد يوفر كل مستشعر كاميرا إشارة تولد دفقًا من البيانات كما اكتشفه المستشعر. يمكن أن تكون هذه المستشعرات واحدة أو أكثر مما يلي:

• مستشعر كاميرا التقريب (14) يطور تدفق كاميرا التقريب (24)؛

• مستشعر كاميرا واسعة (12) يطور تدفق كاميرا واسعة (26)؛

• مستشعر كاميرا محيطة (13) يطور تدفق الكاميرا المحيطي (28).

يمكن إدخال كاميرا التقريب (24) والكاميرا الواسعة (26) والكاميرا المحيطة (26) في تدفقات مستوى البرنامج من المستوى (0) إلى المستوى (4). قد تشمل تدفق التغذية هذه:

• تدفق كاميرا التقريب (24) تتم تغذيته إلى تدفقات مستوى البرنامج 0 و1 (28)، (30)

• تدفق الكاميرا الواسعة (26) تتم تغذيته إلى مستوى البرنامج 2 و3 (32)، (34)

• تدفق الكاميرا المحيطة (28) تتم تغذيته إلى تدفق مستوى البرنامج 4 (36).

تتم معالجة تدفق مستوى البرامج من صفر إلى تدفق مستوى البرنامج 4 بواسطة طبقات التعلم العميق (38، 40، 42، 44 و46). تكون طبقات التعلم العميق هذه (38، 40، 42، 44، 46) متسلسلة (48) ويتم تغذيتها بوحدات نمطية للذاكرة طويلة المدى (LSTM) - T1، T2، T3... Tn (50، 52، 54، 56). يمكن تصنيف المخرجات (58، 60، 62، 64) لوحدات LSTM النمطية من T1 إلى Tn (50، 52، 54، 56) ووضعها في الذاكرة (66، 68، 70 و72) لاستخدامها في المستقبل.

توفر مخرجات LSTM (66، 68، 70، 72) تغذية رجعية من خلال حلقة التغذية الرجعية (74) لضبط آلية الدوران والإمالة التي يتحكم فيها المعالج الحوسبي (23). يتم استخدام تهيئة العدسات (18) و(19) و(20) مع العديد من متغيرات FOV الوسطية التقدمية لتوفير الأطوال البؤرية المطلوبة والخصائص لنظام وطرق الرؤية الذكية. تتضمن متغيرات FOV هذه على سبيل المثال لا الحصر، 360 درجة، 120 درجة، 10 درجة والتي توفر دقة على المستوى البشري تبلغ حوالي 120 بكسل/درجة.

قد تتضمن النماذج الأخرى، على سبيل المثال لا الحصر، متغيرات FOV بمقدار 360 درجة، 120 درجة، 60 درجة، 10 درجة، 2 درجة والتي توفر دقة تفوق المستوى البشري بأكثر من 120 بكسل/درجة.

تعني الإشارة في جميع أنحاء هذه المواصفة إلى "النموذج"، "هذا النموذج"، "النموذج السابق"، "أحد النماذج"، "نموذج"، "نموذج مفضل" "نموذج مفضل آخر" أو لغة مشابهة أنه يتم تضمين السمة المعينة، البنية، أو الخاصية الموصوفة فيما يتعلق بالنموذج في نموذج واحد على الأقل للاختراع الحالي. وبالتالي، فإن ظهور العبارات "في النموذج،"في هذا النموذج"، "في النموذج السابق، في أحد النماذج، في نموذج" ، "في نموذج مفضل"،"في نموذج مفضل آخر"، ولغة مماثلة في جميع أنحاء هذه المواصفة، ولكن ليس بالضرورة، أن تشير جميعها إلى نفس النموذج.

يمكن دمج ملامح، مزايا، وخصائص الاختراع الموصوفة بأي طريقة مناسبة في واحد أو أكثر من النماذج. سيدرك الشخص الماهر في المجال ذي الصلة على أنه يمكن ممارسة الاختراع بدون واحدة أو أكثر من الملامح أو المزايا المحددة لنموذج معين. في حالات أخرى، يمكن إدراك ملامح ومزايا إضافية في نماذج معينة قد لا تكون موجودة في جميع نماذج الاختراع. بينما تم وصف الاختراع الحالي فيما يتعلق ببعض النماذج التوضيحية أو المحددة، يجب أن يكون مفهومًا أن الاختراع لا يقتصر على النماذج التي تم الكشف عنها، ولكن، على العكس من ذلك، يهدف إلى تغطية مختلف التعديلات، البدائل، التعديلات والترتيبات المكافئة كما سيتضح لمن يتمتعون بالمهارات في المجال. الترتيب كما سيكون واضحًا لأولئك المهرة في الفن. يجوز إجراء أي تغييرات، تعديلات، بدائل، مكافئات وما شابه ذلك دون الخروج عن روح ونطاق الكشف.

عناصر الحماية

(1) نظام رؤية ذكي يشتمل على:

أ. كاميرا تم وضعها لإنشاء خرج تدفق بيانات الصورة للكاميرا والتقاط الصور من الكاميرا؛

حيث تكون الكاميرا بزاوية ثابتة قابلة لتشغيلها بمحاذاة مع مرآة مثبتة على وعاء وآلية مشغل آني مائلة قابلة للتشغيل للحصول على الزاوية المرغوبة وعرض للموقع على المرآة والتي يتم التقاطها كصور بواسطة الكاميرا؛

ب. معالج لمعالجة خرج تدفق بيانات الصورة للكاميرا لتوليد واحد أو أكثر من تدفقات بيانات الصورة للكاميرا؛

ج. شبكة عصبية لاستقبال ومعالجة خرج تدفق بيانات الصورة من الكاميرا؛

د. قارئ لقراءة الخرج من الشبكة العصبية للتحكم في موضع الكاميرا؛ و

ه. شبكة تغذية رجعية لمزيد من معالجة خرج تدفق بيانات الصورة للكاميرا للتركيز في النهاية على الهدف موضوع الاهتمام Subject of Interest(SOI).

(2) نظام الرؤية الذكي وفقًا لعنصر الحماية 1، يتضمن كذلك شبكة عصبية للذكاء الاصطناعي وكاميرا ثانية (أو أكثر)، حيث تدير الشبكة العصبية الاصطناعية تسلسلًا هرميًا لزوايا الرؤية المختلفة الناتجة عن الكاميرا الأولى والكاميرا الثانية (أو أكثر).

(3)

3

نظام الرؤية الذكي وفقًا لعنصر الحماية 2، حيث تعمل الكاميرات الأولى والثانية (أو أكثر) كنظام كاميرا واحدة.

4 نظام الرؤية الذكي وفقًا لعنصر الحماية 1، حيث تشتمل الشبكة العصبية للذكاء الاصطناعي على وحدة نمطية تعليمية عميقة تتحكم في الحركة السريعة للكاميرات للعثور على أو اكتشاف أو تتبع واحد أو أكثر من أهداف SOI.

5 نظام الرؤية الذكي وفقًا لعنصر الحماية 1، حيث تقوم الكاميرا بإنتاج تدفق بيانات صورة ذات دقة عادية ويوفر المعالج تدفق بيانات صورة عالية الدقة لهدف واحد أو أكثر، حيث يتم عرض تدفق بيانات الصورة ذات الدقة العادية وتدفق بيانات الصورة عالية الدقة معًا.

6 طريقة لتوفير صور الرؤية الذكية تتضمن استخدام نظام الرؤية الذكي في عنصر الحماية 1، الطريقة المذكورة تضم الخطوات:

أ. وضع كاميرا واحدة على الأقل لإنشاء خرج تدفق بيانات الصورة لكل كاميرا؛

ب. التقاط خرج تدفق بيانات الصورة من كاميرا واحدة على الأقل؛

ج. معالجة خرج تدفق بيانات الصورة من كاميرا واحدة على الأقل لتوليد تدفق بيانات واحد أو أكثر؛

د. مزيد من المعالجة لخرج تدفق بيانات الصورة من كاميرا واحدة على الأقل عبر شبكة عصبية؛

ه. قراءة الخرج من الشبكة العصبية للتحكم في موضع كاميرا واحدة على الأقل؛ و

و. مزيد من المعالجة لدفق بيانات الصورة لكاميرا واحدة على الأقل من خلال شبكة تغذية رجعية.

7 الطريقة وفقًا لعنصر الحماية 6، تموضع (تحديد موضع) بشكل إضافي لكاميرا ثانية (أو أكثر) لإنشاء خرج تدفق بيانات الصورة الثانية، وشبكة عصبية اصطناعية تدير تسلسلًا هرميًا لزوايا عرض مختلفة تم إنشاؤها بواسطة الكاميرا الأولى والثانية (أو أكثر) كاميرا.

8 الطريقة وفقًا لعنصر الحماية 7، حيث تعمل الكاميرات الأولى والثانية (أو أكثر) كنظام كاميرا واحدة.

9

الطريقة وفقًا لعنصر الحماية 6، حيث تتضمن المعالجة الإضافية لتدفق بيانات الصورة لكاميرا واحدة على الأقل من خلال شبكة تغذية رجعية معالجة تدفق بيانات الصورة لكاميرا واحدة على الأقل من خلال وحدة تعلم عميق تتحكم في الحركة السريعة للكاميرا للعثور أو اكتشاف أو تتبع هدف واحد أو أكثر من أهداف موضوع الاهتمام SOI.

10 الطريقة وفقًا لعنصر الحماية 6، حيث تؤدي معالجة خرج تدفق بيانات الصورة لكاميرا واحدة على الأقل لتوليد واحد أو أكثر من تدفقات البيانات إلى تدفق بيانات صورة ذات دقة عادية ومعالجة إضافية لخرج تدفق بيانات الصورة من كاميرا واحدة على الأقل من خلال شبكة عصبية توفر تدفق بيانات صورة عالية الدقة لواحد أو أكثر من أهداف SOI، حيث يتم عرض تدفق بيانات الصورة ذات الدقة العادية وتدفق بيانات الصورة عالية الدقة معًا.

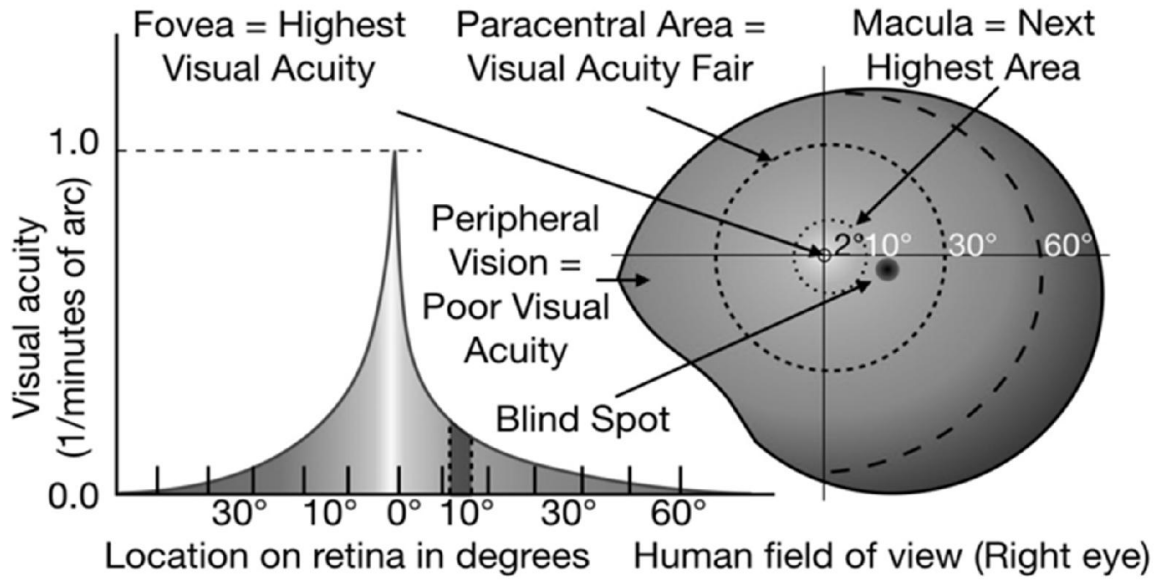


Figure 1



Figure 2

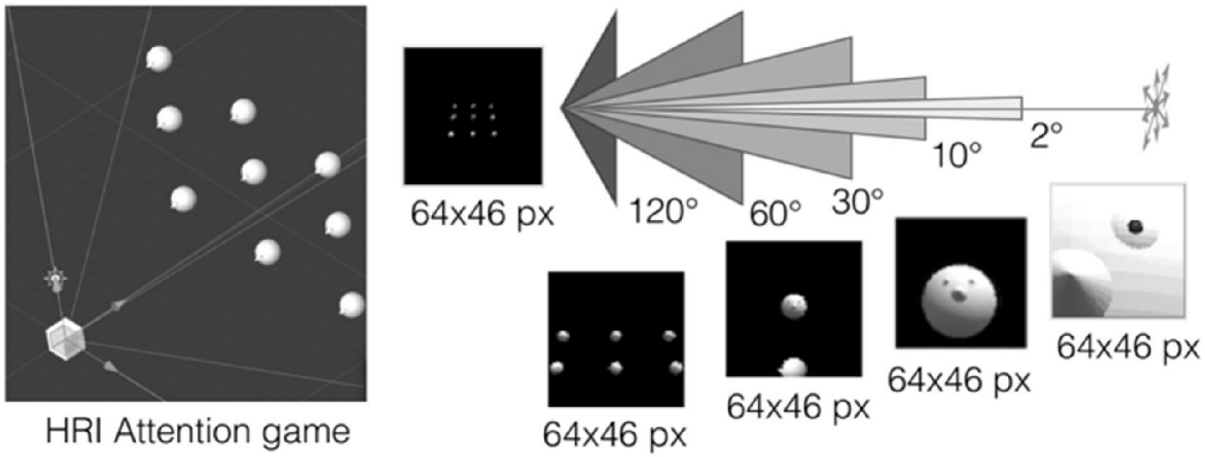

HRI Attention game
64x46 px
120°
60°
30°
10°
2°
64x46 px
64x46 px
64x46 px
64x46 px

Figure 3

Experiment	Cameras	FOV	HRI Game (PPO-Attention)	MNIST (RAM)
Baseline			0.24	
1	1	10	-0.73	0.680
2	1	30	-0.47	0.900
3	1	60	-0.60	0.935
4	2	60/10	-0.59	0.905
5	2	60/30	-0.25	0.950
6	2	30/10	-0.16	0.865
7	3	60/30/10	0.04	0.970

Figure 4 (a)

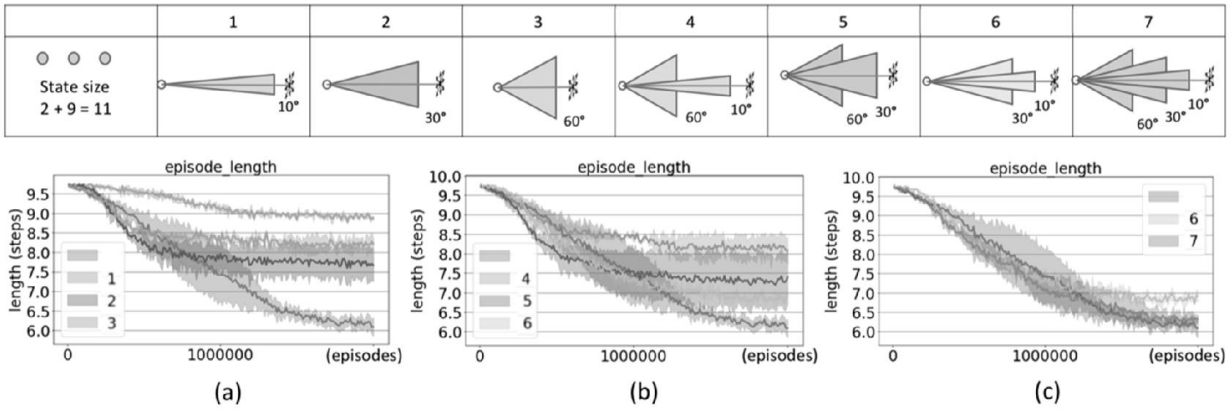


Figure 4 (b) Figure 4 (c) Figure 4 (d)

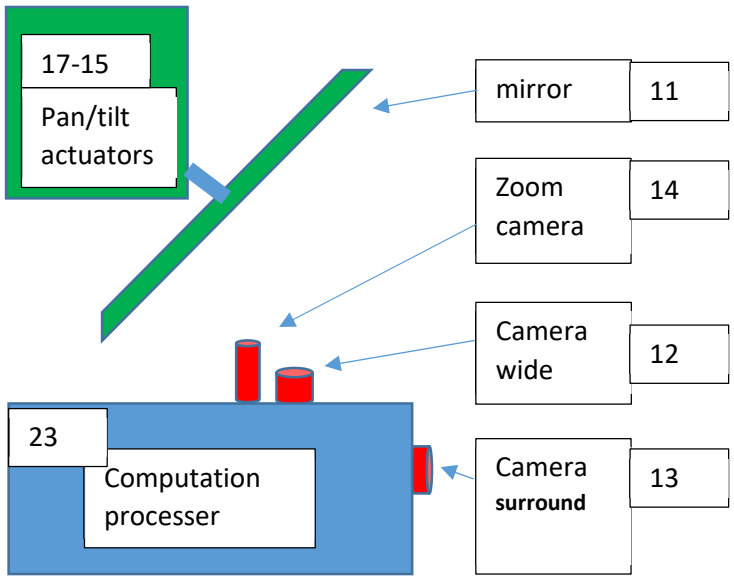

17-15
Pan/tilt actuators
mirror
11
Zoom camera
14
Camera wide
12
23
Computation processer
Camera surround
13

Figure 1

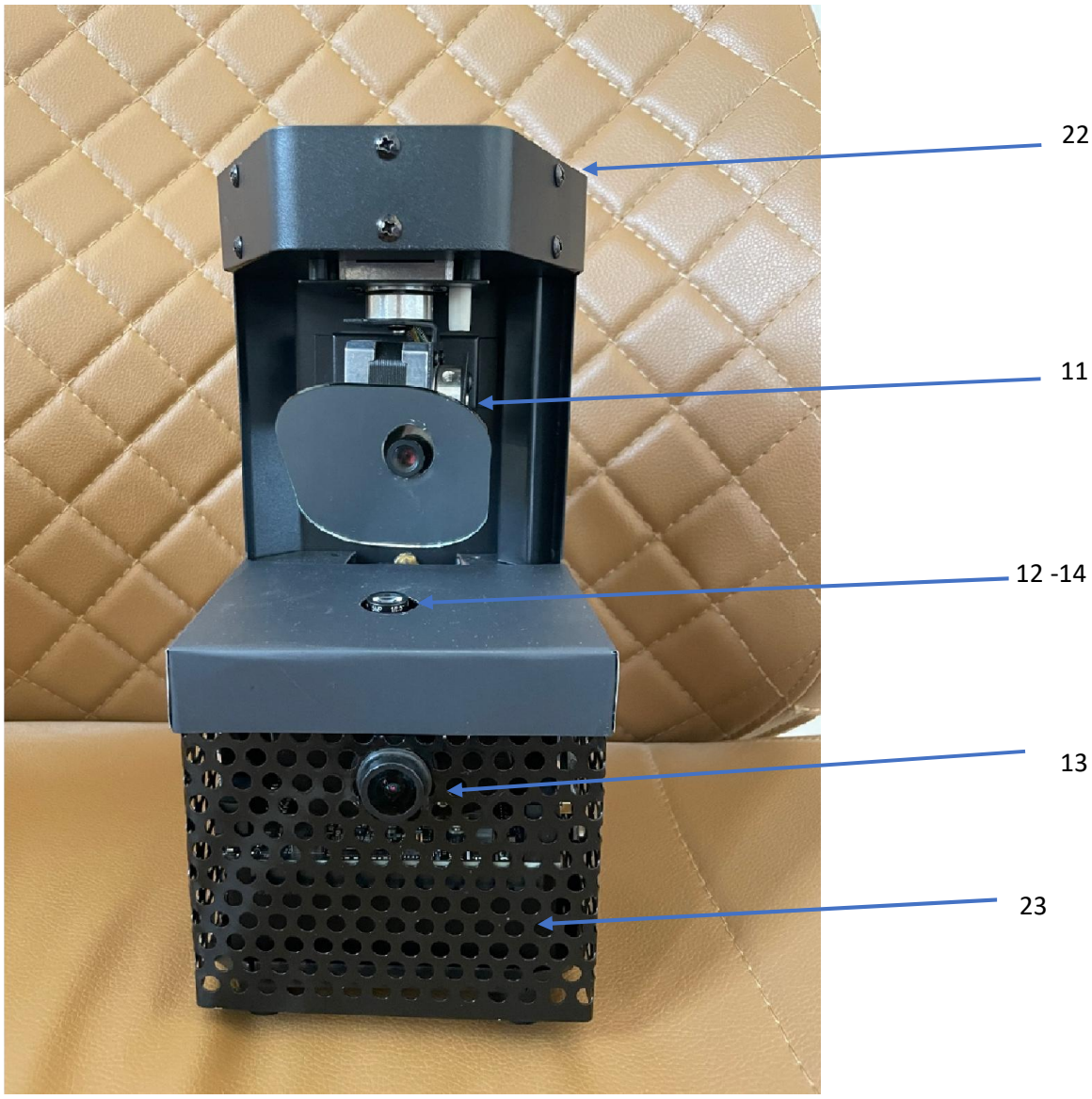


Figure 2

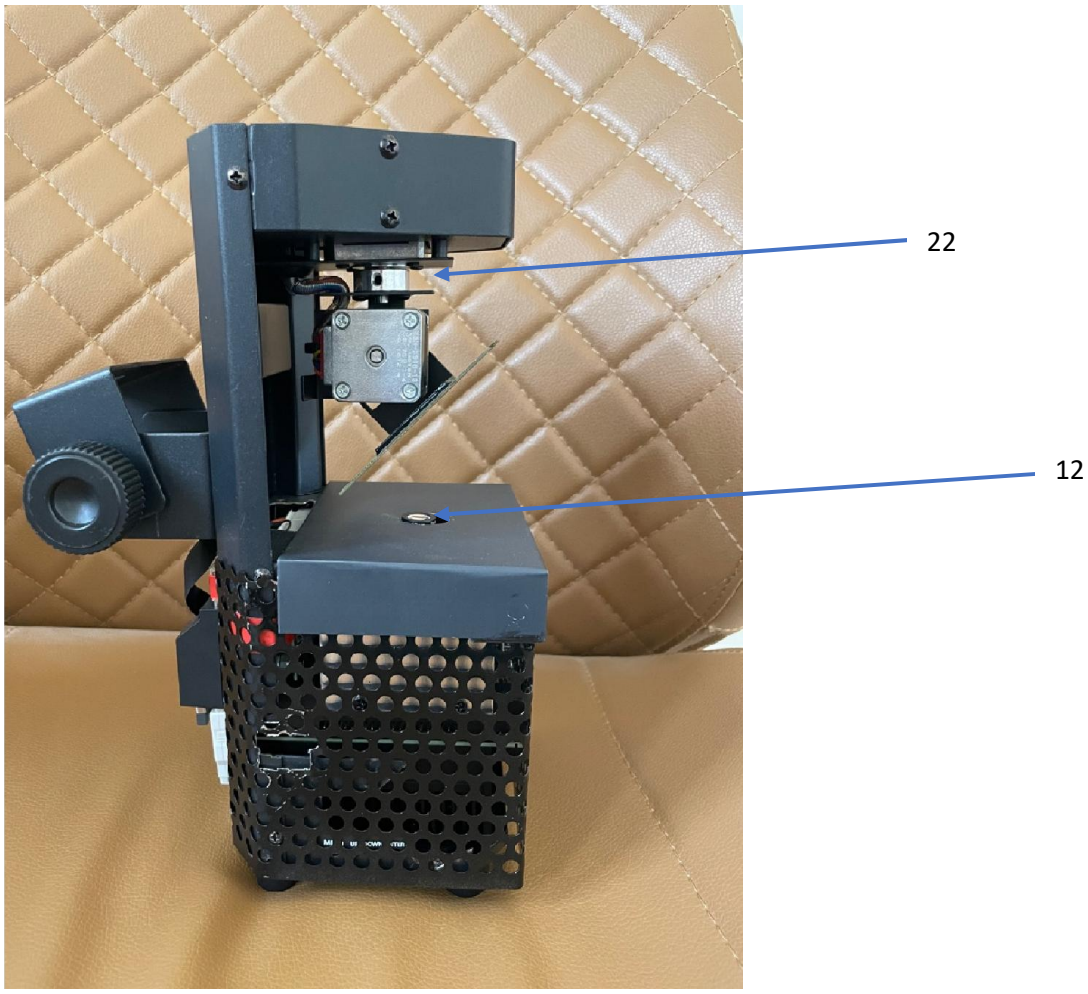


Figure 7

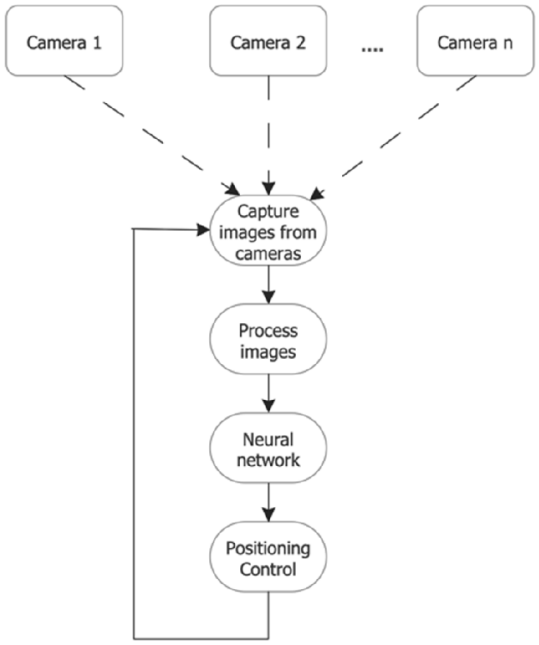


Figure 8

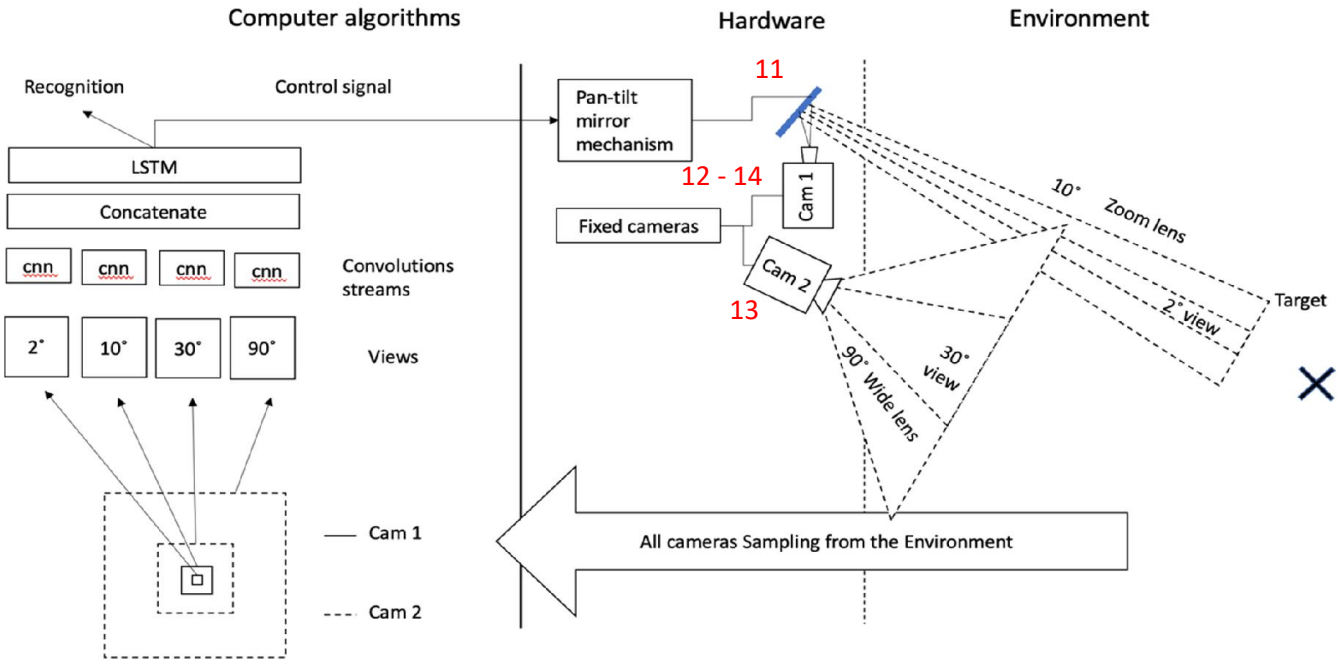


Figure 9

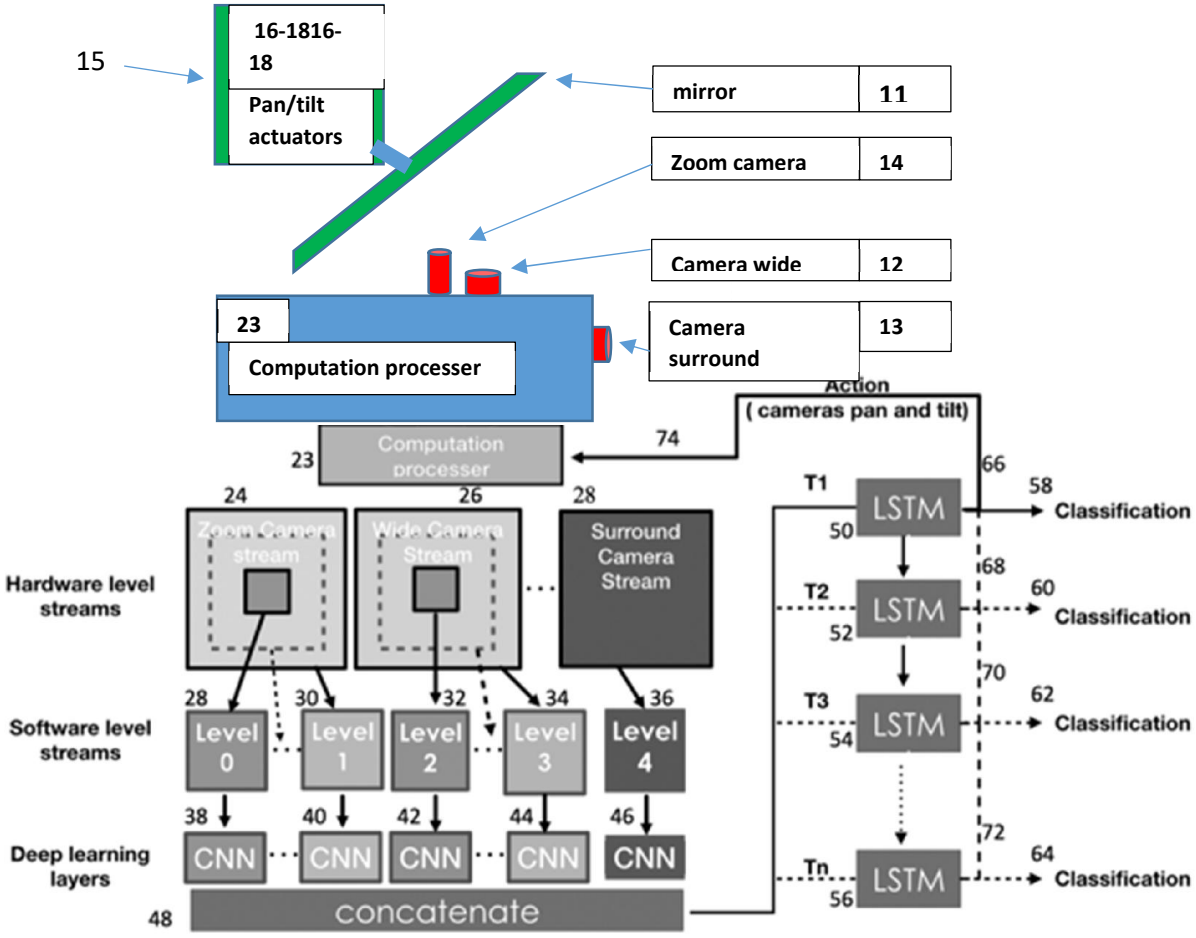
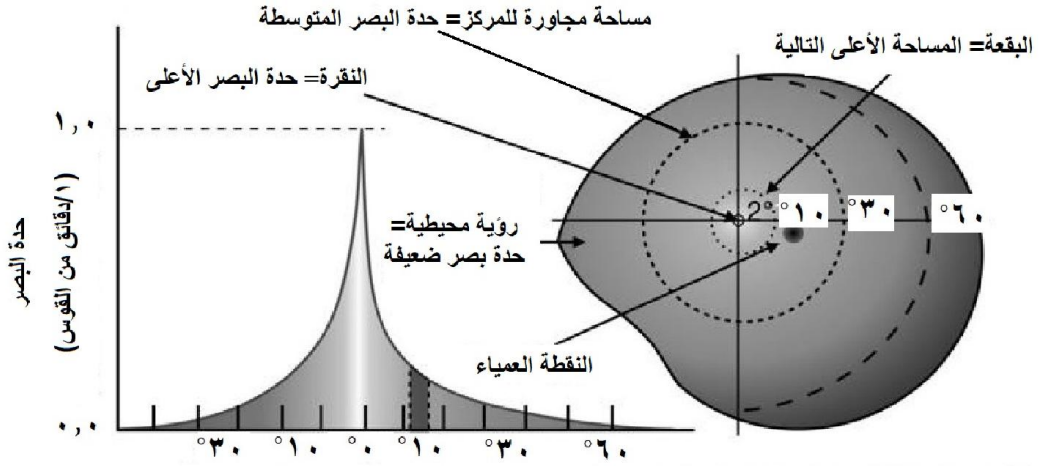
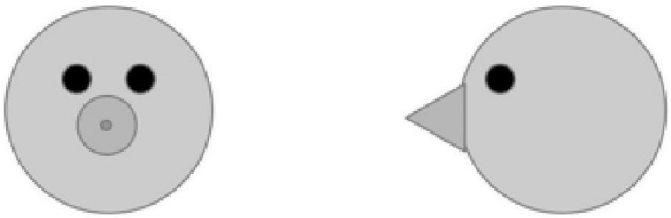

15
16-1816-18
Pan/tilt actuators
mirror
11
Zoom camera
14
Camera wide
12
23
Computation processer
Camera surround
13
Action
(cameras pan and tilt)
74
23
Computation processer
66
24
26
28
Zoom Camera stream
Wide Camera Stream
Surround Camera Stream
Hardware level streams
T1
LSTM
50
58
Classification
T2
LSTM
52
68
60
Classification
T3
LSTM
54
70
62
Classification
Tn
LSTM
56
72
64
Classification
28
30
32
34
36
Level 0
Level 1
Level 2
Level 3
Level 4
Software level streams
38
40
42
44
46
CNN
CNN
CNN
CNN
CNN
Deep learning layers
48
concatenate

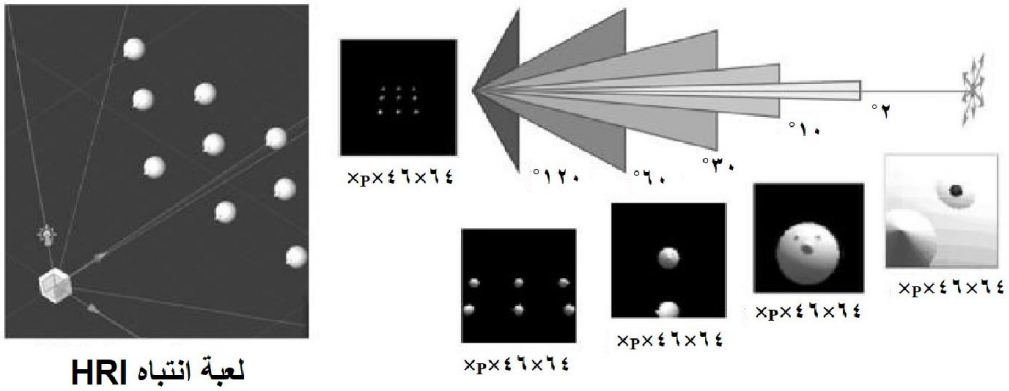
Figure 10


مساحة مجاورة للمركز= حدة البصر المتوسطة
البقعة= المساحة الأعلى التالية
النقرة= حدة البصر الأعلى
١,٠
٠,٠
حدة البصر
(١/دقائق من القوس)
رؤية محيطية=
حدة بصر ضعيفة
النقطة العمياء
٢°
١٠°
٣٠°
٦٠°
٣٠° ١٠° ٠° ١٠° ٣٠° ٦٠°
المجال البشري للرؤية (العين اليمنى)
الموقع على الشبكية بالدرجات

شكل ١



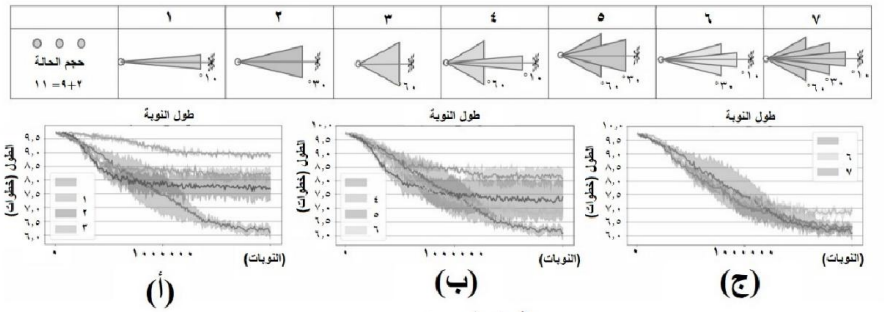
شكل ٢



شكل ٣

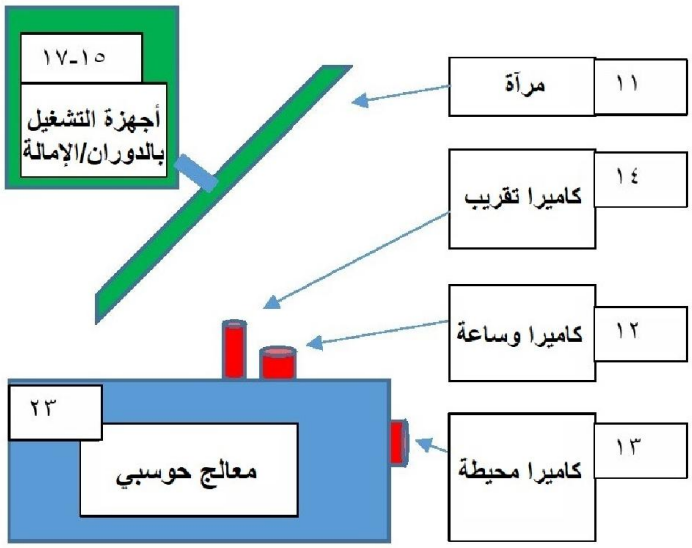
التجربة	الكاميرات	FOV	لعبة HRI (انتباه PPO)	MNIST (RAM)
القيمة الأساسية			٠,٢٤	
١	١	١٠	٠,٧٣ -	٠,٦٨٠
٢	١	٣٠	٠,٤٧ -	٠,٩٠٠
٣	١	٦٠	٠,٦٠ -	٠,٩٣٥
٤	٢	٦٠/١٠	٠,٥٩ -	٠,٩٠٥
٥	٢	٦٠/٣٠	٠,٢٥ -	٠,٩٥٠
٦	٢	٣٠/١٠	٠,١٦ -	٠,٨٦٥
٧	٣	٦٠/٣٠/١٠	٠,٠٤	٠,٩٧٠

شكل ٤أ

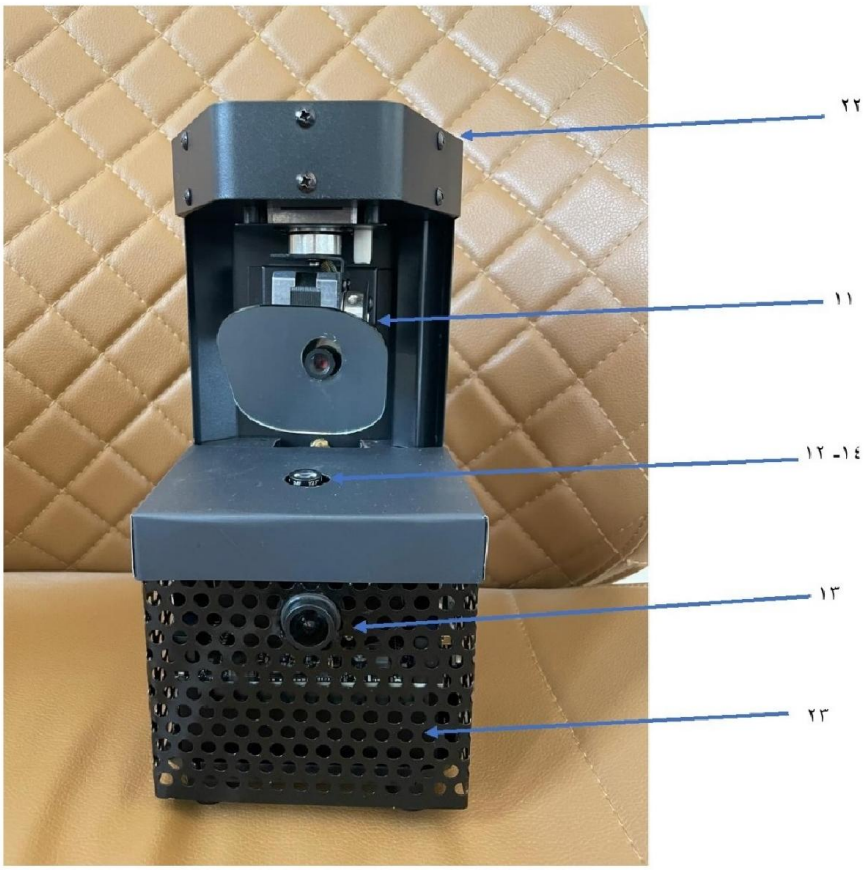


(أ) (ب) (ج)

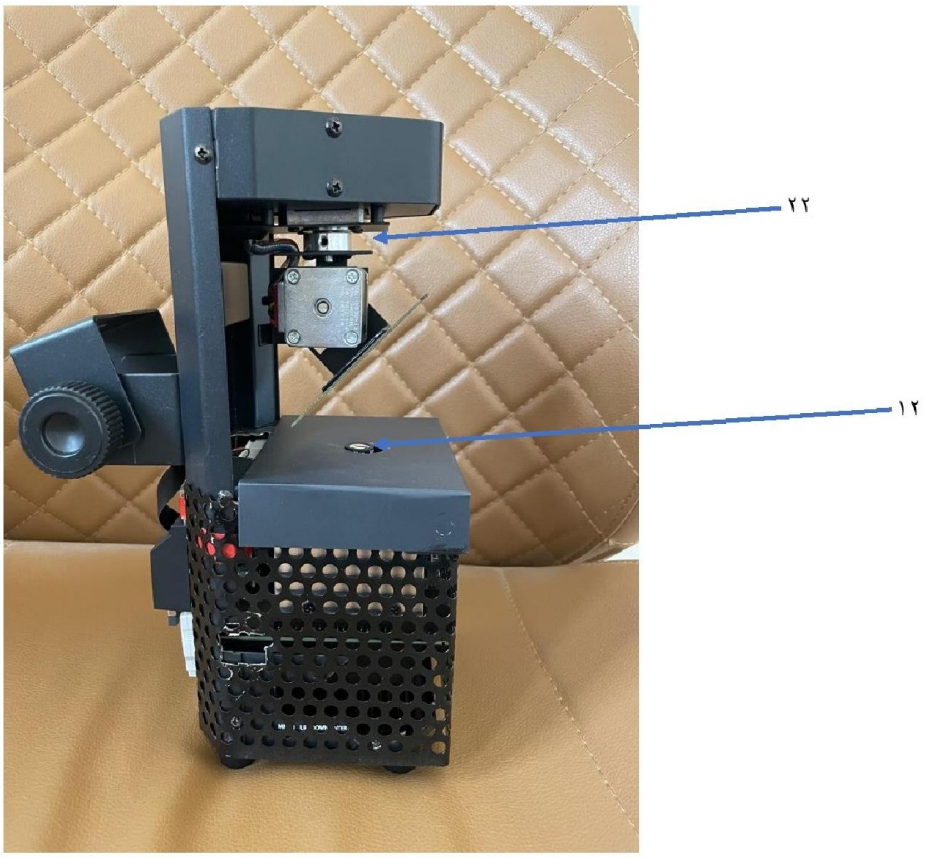
شكل ٤ب شكل ٤ج شكل ٤د


١٥-١٧
أجهزة التشغيل بالدوران/الإمالة
مرآة
١١
كاميرا تقريب
١٤
كاميرا وساعة
١٢
٢٣
معالج حوسبي
كاميرا محيطة
١٣

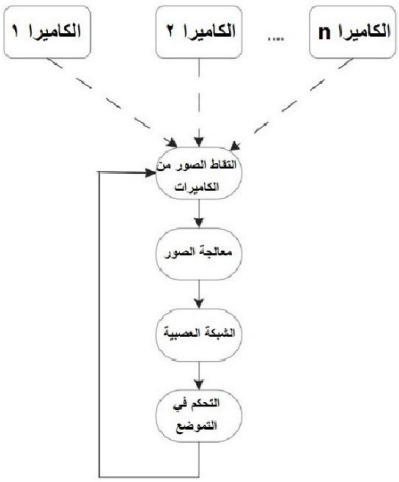
شكل ٥



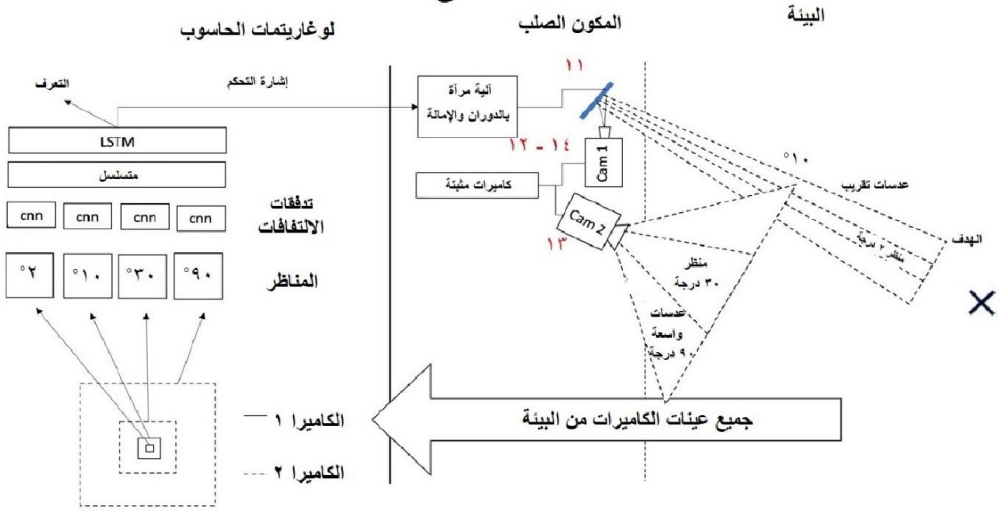
شكل ٦



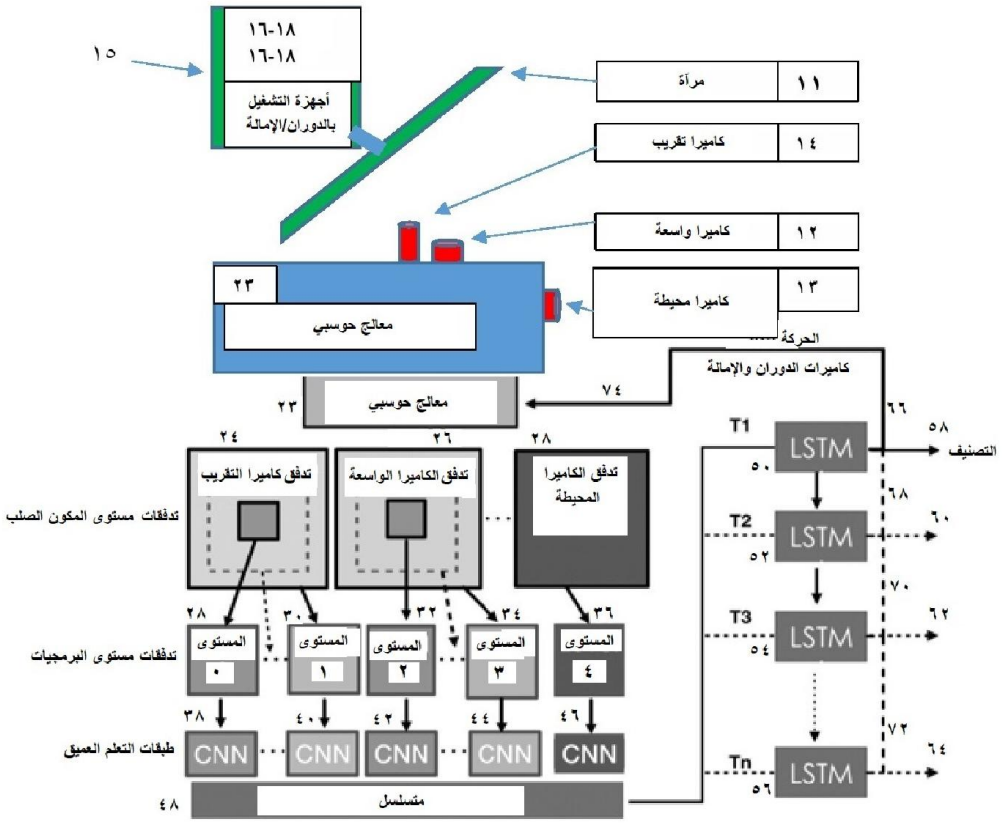
شکل ۷


الكاميرا ١
الكاميرا ٢
....
الكاميرا n
التقاط الصور من الكاميرات
معالجة الصور
الشبكة العصبية
التحكم في التموضع

شكل ٨


لوغاريتمات الحاسوب
المكون الصلب
البيئة
التعرف
إشارة التحكم
LSTM
متسلسل
cnn
cnn
cnn
cnn
تدفقات الالتفافات
°٢
°١٠
°٣٠
°٩٠
المناظر
الكاميرا ١
الكاميرا ٢
آلية مراة بالدوران والإمالة
١١
١٢ - ١٤
كاميرات مثبتة
Cam 1
Cam 2
١٣
°١٠
عدسات تقريب
الهدف
منظر ٣٠ درجة
عدسات واسعة ٩٠ درجة
جميع عينات الكاميرات من البيئة

شكل ٩


١٥
١٦-١٨
١٦-١٨
أجهزة التشغيل بالدوران/الإمالة
مرآة
١١
كاميرا تقريب
١٤
كاميرا واسعة
١٢
كاميرا محيطة
١٣
٢٣
معالج حوسبي
الحركة
كاميرات الدوران والإمالة
٧٤
٢٣
معالج حوسبي
٢٤
٢٦
٢٨
تدفق كاميرا التقريب
تدفق الكاميرا الواسعة
تدفق الكاميرا المحيطة
تدفقات مستوى المكون الصلب
٢٨
٣٠
٣٢
٣٤
٣٦
المستوى ٠
المستوى ١
المستوى ٢
المستوى ٣
المستوى ٤
تدفقات مستوى البرمجيات
٣٨
٤٠
٤٢
٤٤
٤٦
CNN
CNN
CNN
CNN
CNN
طبقات التعلم العميق
٤٨
متسلسل
٦٦
٥٨
التصنيف
T1
٥٠
LSTM
٦٨
٦٠
T2
٥٢
LSTM
٧٠
٦٢
T3
٥٤
LSTM
٧٢
٦٤
Tn
٥٦
LSTM

شكل ١٠