1 Acknowledgement

I would like to express my gratitude towards my guide Prof. A.M.Bhadgale of Computer Engineering Department, who has been very concerned and have adided for all the help essentials for the preparation of this work. He has helped me to explore this vast topic in an organised manner and provided me with all the ideas on how to work towards a research oriented venture.

I am thankful to Prof.M.V.Marathe, Head of Department, Computer Engineering, for the motivation and inspiration that triggered me for the seminar work.

Shantanu Kalamdane T150074264

(T.E. Computer Engineering)

2 Abstract

One of the goals of Artificial intelligence (AI) is the realization of natural dialogue between humans and machines. In recent years, the dialogue systems, also known as interactive conversational systems are the fastest growing area in AI. Many companies have used the dialogue systems technology to establish various kinds of Virtual Personal Assistants(VPAs) based on their applications and areas, such as Microsofts Cortana, Apples Siri, Amazon Alexa, Google Assistant, and Facebooks M. However, in this proposal, we have used the multi-modal dialogue systems which process two or more combined user input modes, such as speech, image, video, touch, manual gestures, gaze, and head and body movement in order to design the Next- Generation of VPAs model. The new model of VPAs will be used to increase the interaction between humans and the machines by using different technologies, such as gesture recognition, image/video recognition, speech recognition, the vast dialogue and conversational knowledge base, and the general knowledge base. Moreover, the new VPAs system can be used in other different areas of applications, including education assistance, medical assistance, robotics and vehicles, disabilities systems, home automation, and security access control.

Contents

3 Introduction

Spoken dialogue systems are intelligent agents that are able to help users finish tasks more efficiently via spoken interactions. Also, spoken dialogue systems are being incorporated into various devices such as smart-phones, smart TVs, in car navigating system. Also, Dialogue systems or conversational systems can support a wide range of applications in business enterprises, education, government, healthcare, and entertainment. Personal assistants, known by various names such as virtual personal assistants, intelligent personal assistants, digital personal assistants, mobile assistants, or voice assistants. Many companies have used the spoken dialogue systems to design their dialogue system device, such as Microsofts Cortana, Apples Siri, Amazon Alexa, Google Assistant, Samsung S Voice, Nuance Dragon, and Facebooks M. These companies used different approaches to design and improve their dialogue systems. There are many techniques used to design the VPAs, based on the application and its complexity. For example, Google has improved the Google Assistant by using the Deep Neural Networks (DNN) method which highlights the main components of dialogue systems and new deep learning architectures used for these components. Also, Microsoft used the Microsoft Azure Machine Learning Studio with other Azure components to improve the Cortana dialogue system.

In this proposal, we propose an approach that will be used to design the Next-Generation of Virtual Personal Assistants, increasing the interaction between users and the computers by using the Multi-modal dialogue system with techniques including the gesture recognition, image/video recognition, speech recognition, the vast dialogue and conversational knowledge base, and the general knowledge base. Moreover, our approach will be used in different tasks including education assistance, medical assistance, robotics and vehicles, disabilities systems, home automation, and security access control. To design the Next-Generation of Virtual Personal Assistants with high accuracy, we added some components to the original structure of general dialogue systems to change the general model to Multi-modal dialogue systems, such as ASR Model, Gesture Model , Graph Model, Interaction Model, User Model, Input Model, Output Model, Inference Engine, Cloud Servers and Knowledge Base.

4 Proposed Architecture

In this proposal, we have used the multi-modal dialogue systems which process two or more combined user input modes, such as speech, image, video, touch, manual gestures, gaze, and head and body movement in order to design the Next-Generation of VPAs model. We have modified and added some components in the original structure of general dialogue systems, such as ASR Model, Gesture Model, Graph Model, Interaction Model, User Model, Input Model, Output Model, Inference Engine, Cloud Servers and Knowledge Base. The following is the structure of the Next-Generation of Virtual Personal Assistants:

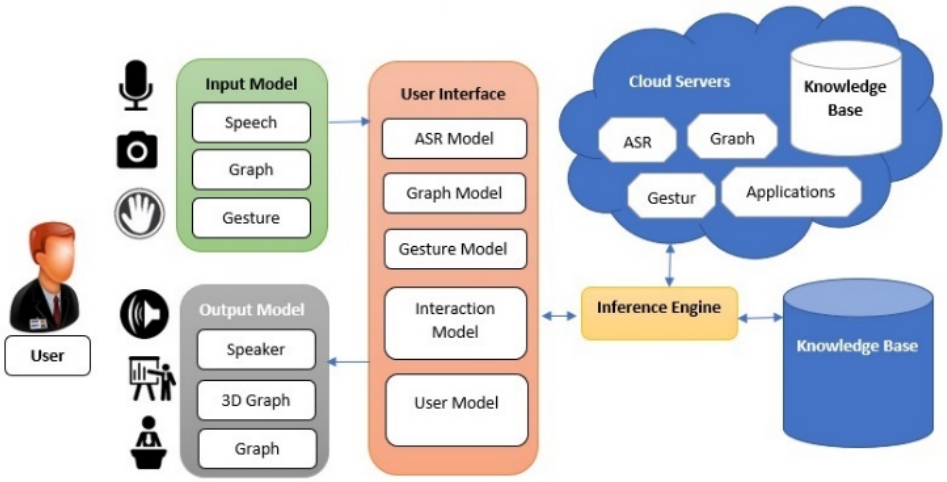

Figure 1: The Structure of The Next-Generation of Virtual Personal Assistants

4.1 Knowledge Base

There are two knowledge bases. The first is the online and the second is local knowledge base which include all data and facts based on each model, such as facial and body data sets for gesture modal, speech recognition knowledge bases, dictionary and spoken dialog knowledge base for ASR modal, video and image body data sets for Graph Model, and some users information and the setting system.

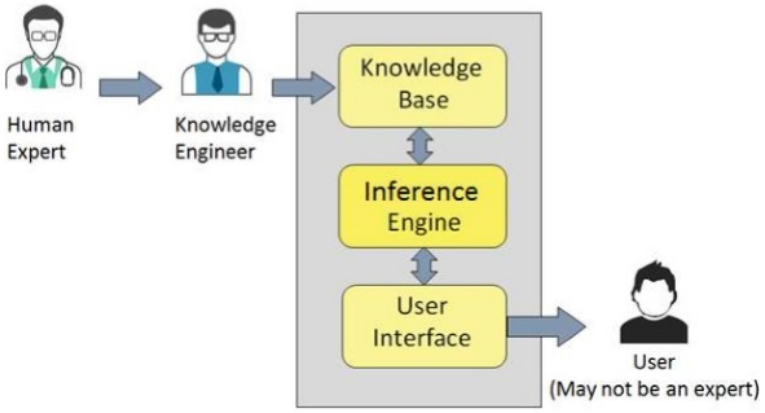

Figure 2: The Knowledge Base

4.1.1 Knowledge System

A Knowledge-Based System (KBS) is a computer program that reasons and uses a knowledge base to solve complex problems. The one common theme that unites all knowledge based systems is an attempt to represent knowledge explicitly and a reasoning system that allows it to derive new knowledge.

4.1.2 Interference Engine

An inference engine is a computer program that applies artificial intelligence to try to obtain answers or responses to queries from a knowledge base. A programs protocol for navigating through the rules and data in a knowledge system in order to solve the problem. The major task of the inference engine is to select and then apply the most appropriate rule at each step as the expert system runs, which is called rule-based reasoning. The part of a decision support system that performs the reasoning function.

4.2 Graph Model

The Graph Model analyzes video and image in real-time by using the Graph Model and extracts frames of the video that collect by the camera and the input model; then it sends those frames and images to the Graph Model and applications in Cloud Servers for analyzing those frames and images and returning the result.

4.3 Gesture Model

The gesture model uses the camera and Kinect in the input model to read the movements of the human body and the facial expressions; then it sends all data to the gesture model and applications in Cloud Servers to analyze those frames and images and returning the result.

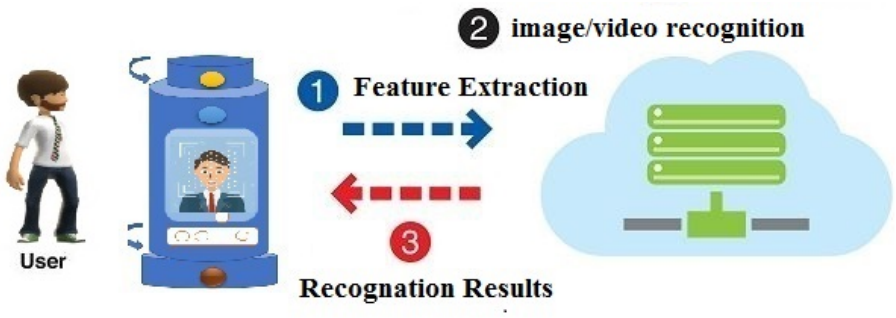

Figure 3: The Graph Model

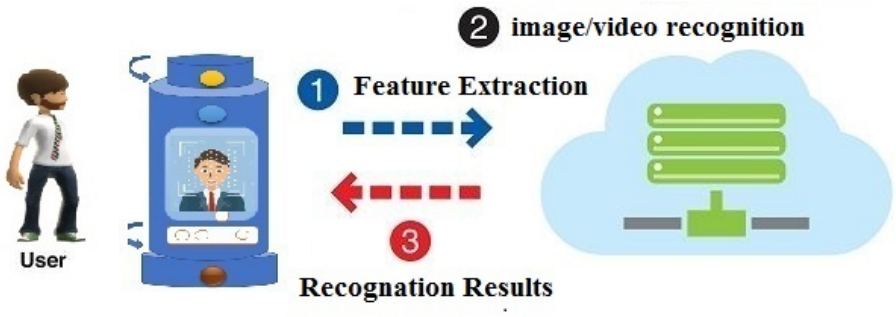

Figure 4: The Graph Model

4.4 ASR Model

The speech recognition model will work in real-time with the microphone in the input model with the ASR model in Cloud Servers to recognize the utterances that a user speaks into a microphone and then convert it to text; then it sends the text to the applications in Cloud Servers to analyze the text and returning the result.

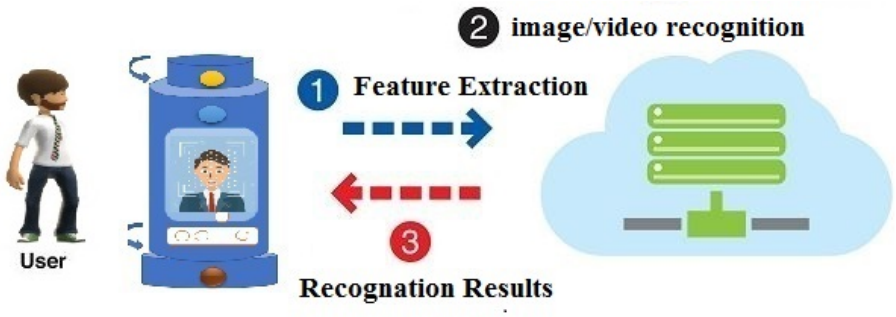

Figure 5: The Graph Model

4.5 Interaction Model

This is the main model that will be used to provide interaction between users of the system and the system models by receiving the data from the input model and analyzing the data to send for each model based on its tasks, then returning result that will be used to make the final decision.

4.6 Interference Engine

The inference engine works together with the Interaction Model in the chain of conditions and derivations and finally deduces the outcome. they analyze all the facts and rules, then sorts them before concluding to a solution.

4.7 User Model

This model has all information about the users that will use the system. It can include personal information such as users' names and ages, their interests, their skills and knowledge, their goals and plans, their preferences and their dislikes or data about their behavior and their interactions with the system. All information will be collected by asking the user some questions then storing all answers in the Knowledge Base.

User models are the subdivision of human computer interaction which describes the process of building up and modifying a conceptual understanding of the user. The main goal of user models is customization and adaptation of systems to the user's specific needs. The system needs to "say the 'right' thing at the 'right' time in the 'right' way".To do so it needs an internal representation of the user. Another common purpose is modeling specific kinds of users, including modeling of their skills and declarative knowledge, for use in automatic software-tests.User-models can thus serve as a cheaper alternative to user testing.

4.8 Input Model

This model has all information about the users that will use the system. It can include personal information such as users' names and ages, their interests, their skills and knowledge, their goals and plans, their preferences and their dislikes or data about their behavior and their interactions with the system. All information will be collected by asking the user some questions then storing all answers in the Knowledge Base.

4.9 Output Model

This model will receive the final decision from the Interaction Model with an explanation, then it will choose the perfect output device to show the result such data show, speakers or screen based on the result.

5 Graph Model

5.1 Computer Vision

Computer vision is an interdisciplinary field that deals with how computers can be made to gain high-level understanding from digital images or videos. From the perspective of engineering, it seeks to automate tasks that the human visual system can do. It is concerned with the automatic extraction, analysis and understanding of useful information from a single image or a sequence of images. It involves the development of a theoretical and algorithmic basis to achieve automatic visual understanding. The image data can take many forms, such as video sequences, views from multiple cameras, or multi-dimensional data from a medical scanner.

5.2 Image Recognition

Among different types, Computer Vision is a major flavor of airtificial intelligence. It refers to the ability of computers to acquire, process, and analyze data that is coming primarily from visual sources. The ability to track or predict movement for instance, but could also include data from heat sensors and other similar source. You might call image recognition a subset of computer vision, in that it refers to the ability of a computer to see, to decipher and understand the information fed to it from an image, be it a still, video, graphic, or even live. This is no small feat.

Consider that a photo, image, or video is infinitely more complex and open-ended than the words that make up a sentence. Think of a newborn that is dazzled by light and color, and you begin to a touch the experience of a computer that has no pre-defined way of understanding what all the various data in an image are. In fact, to a computer, a photo is simply a bunch of tiny colored dots arrayed in pattern (what we call pixels, to be more precise). In order to make sense of what those dots all mean, the computer needs to first understand that patterns that make up things called objects, and objects exist in space and have dimensions, and on an on. Thats a pretty steep learning curve. (In fact, as humans we use about half our brain power to process visual information!)

A digital image represents a matrix of numerical values. These values represent the data associated with the pixel of the image. The intensity of the different pixels, averages to a single value, representing itself in a matrix format. The information fed to the recognition systems is the intensities and the location of different pixels in the image. With the help of this information, the systems learn to map out a relationship or pattern in the subsequent images supplied to it as a part of the learning process. After the completion of the training process, the system performance on test data is validated. In order to improve the accuracy of the system to recognize images, intermittent weights to the neural networks are modified to improve the accuracy of the systems. Some of the algorithms used in image recognition are SIFT (Scale-invariant Feature Transform), SURF (Speeded Up Robust Features), PCA (Principal Component Analysis), and LDA (Linear Discriminant Analysis).

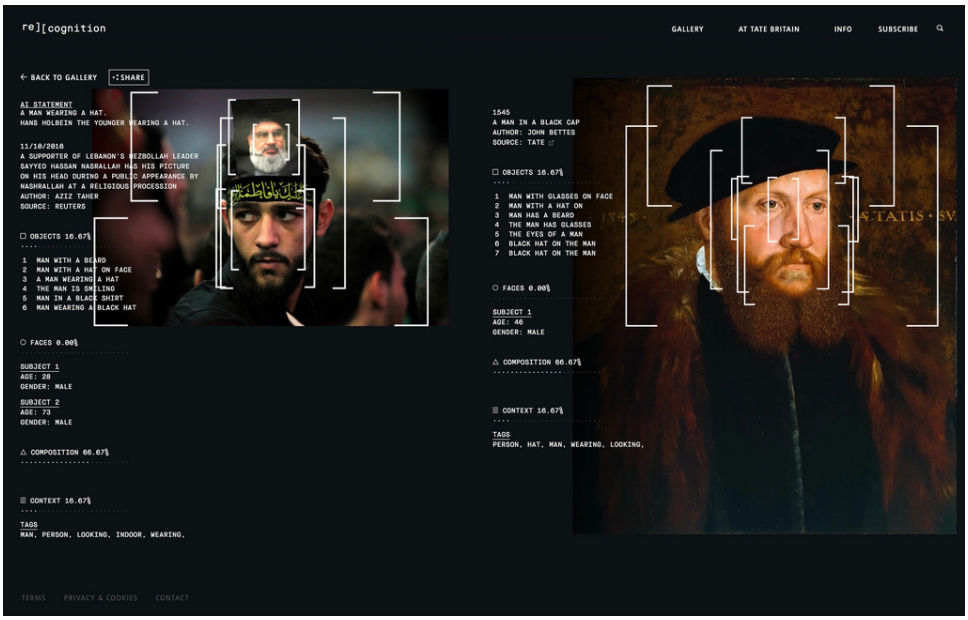

Figure 6: small sections of an image feeding out data

6 Gesture Model

6.1 Gesture Recognition

Gesture recognition is a topic in computer science and language technology with the goal of interpreting human gestures via mathematical algorithms. Gestures can originate from any bodily motion or state but commonly originate from the face or hand. Users can use simple gestures to control or interact with devices without physically touching them. Many approaches have been made using cameras and computer vision algorithms to interpret sign language. However, the identification and recognition of posture, gait, proxemics, and human behaviors is also the subject of gesture recognition techniques. Gesture recognition is a way for computers to begin to understand human body language, thus building a richer bridge between machines and humans than primitive text user interfaces or even GUIs (graphical user interfaces), which still limit the majority of input to keyboard and mouse and interact naturally without any mechanical devices. Using the concept of gesture recognition, it is possible to point a finger at this point will move accordingly. This could make conventional input on devices such and even redundant.

6.2 Gesture recognition features:

- More Accurate
- High Stability
- Time saving to unlock a device

6.3 Major application areas of gesture recognition in the current scenario are:

- Automotive Sector
- Consumer Electronics sector
- Transit Sector
- Defense

7 Kinect

KINECT 3D sensing camera (development code name "Project Natal") is used in the hardware part, while the function of real-time capture, microphone input, speech recognition, image recognition, interactive community features and so on are also included in this part to accurately identify human body and real time capture. Kinect has three lenses, left and right sides of the lens respectively are the IR emitter and infrared sensors that can be used for position control and collecting in-depth data (the distance from the camera) by matching.The middle one is RGB color cameras which is used for collecting image location to locate the image positions. Color camera support 1280*960 resolution imaging, infrared camera support max 640*480 imaging.

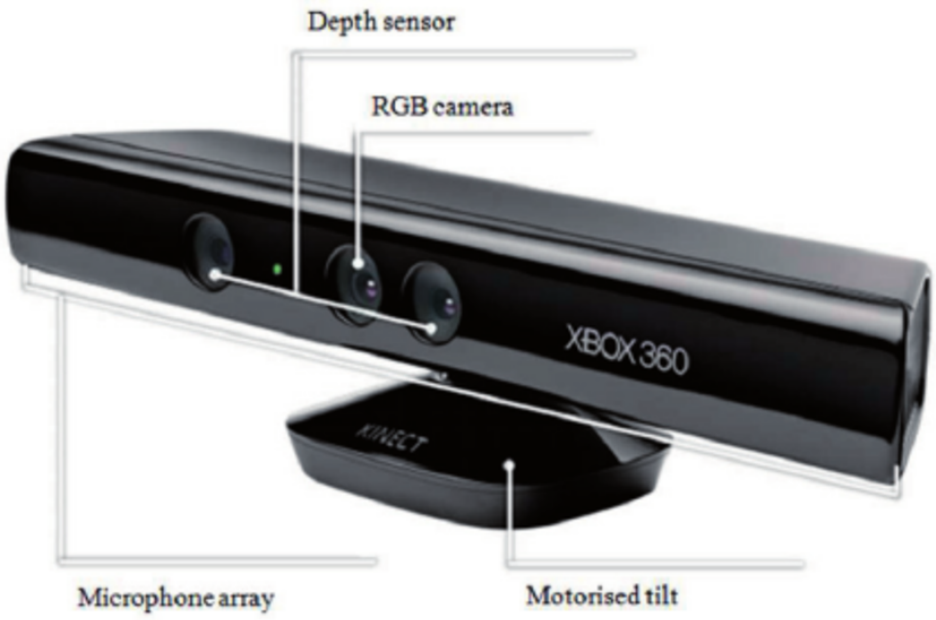

Figure 7: Kinect Devices

8 Object Detection by Voila Jones

The object detection procedure classifies images based on the value of simple features. There are many motivations for using features rather than the pixels directly. The most common reason is that features can act to encode ad-hoc domain knowledge that is difficult to learn using a finite quantity of training data. For this system there is also a second critical motivation for features: the feature based system operates much faster than a pixel-based system. The simple features used is a recollection of Haar basis functions which have been used earlier. More specifically, we use three kinds of features. "The value of a two-rectangle feature is the difference between the sum of the pixels within two rectangular regions". The regions have the same size and shape and are horizontally or vertically adjacent. A three-rectangle feature computes the sum within two outside rectangles subtracted from the sum in a center rectangle. A four-rectangle feature computes the difference between diagonal pairs of rectangles. Given that the base resolution of the detector is 24x24, the exhaustive set of rectangle features is quite large, over 180,000 . Compared to the Haar basis, the set of rectan- gle features proves out to be efficient.

8.0.1 Integral Image

Rectangle features can be calculated very rapidly using an intermediate representation of the image which is known as the integral image. The integral image, contains the sum of the pixels above and to the left of, inclusive:

$ii(x, y) = \sum_{x^1 \leq x, y^1 \leq y} i(x^1, y^1)$

Given example images (x1,y1),...(xn,yn) where yi = 0 for negative and positive examples respectively. Initialize weights $w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ for yi = 0,1 respectively, where l and m are the numbers of negatives and positives respectively. For t = 1,...,T Normalize the weights,

$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n}} w_{t,j}$

so that $w_t.$ is a probability distribution.

For each feature, train a classifier which is restricted to using a single feature. Error evaluated with respect to, $w_t, \epsilon_j = \sum_i w_i |h_j(x_i) - y_i|$. Choose the classifier, h_t , with the lowest error ϵ_t. And then update the weights:

$w_t + 1, i = w_{t,i} \beta^{1-e_i}$

where $e_i = 0$, if example x_i is classified correctly, $e_i = 1$ otherwise and $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$

The final strong classifier is:

h(x) = For x = 1 , $\sum_{t=1}^{T}\alpha_t h_t(x) \geq 1/2\sum_{t=1}^{T}\alpha_t$

For x = 0 , Otherwise

where $\alpha_t = log\frac{1}{\beta_t}$

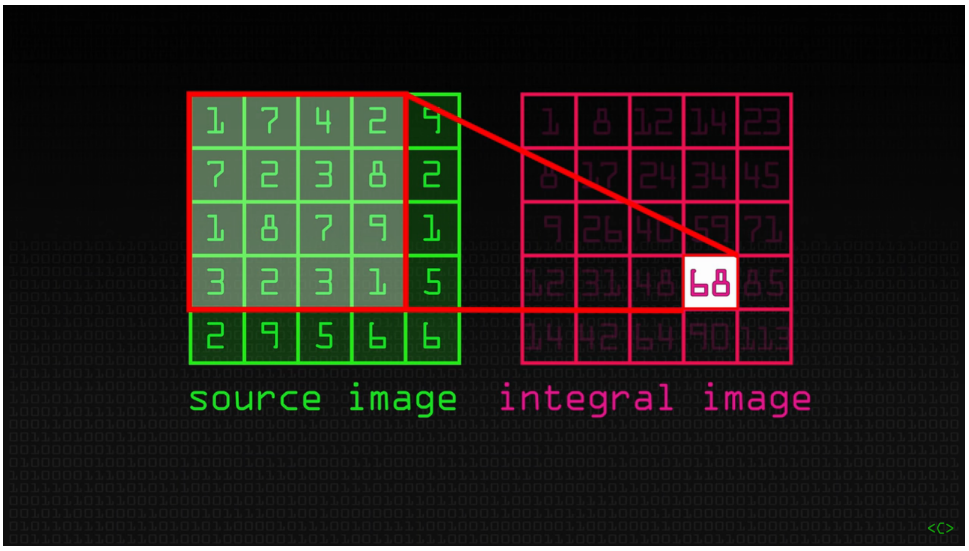

Figure 8: Source and Integral image

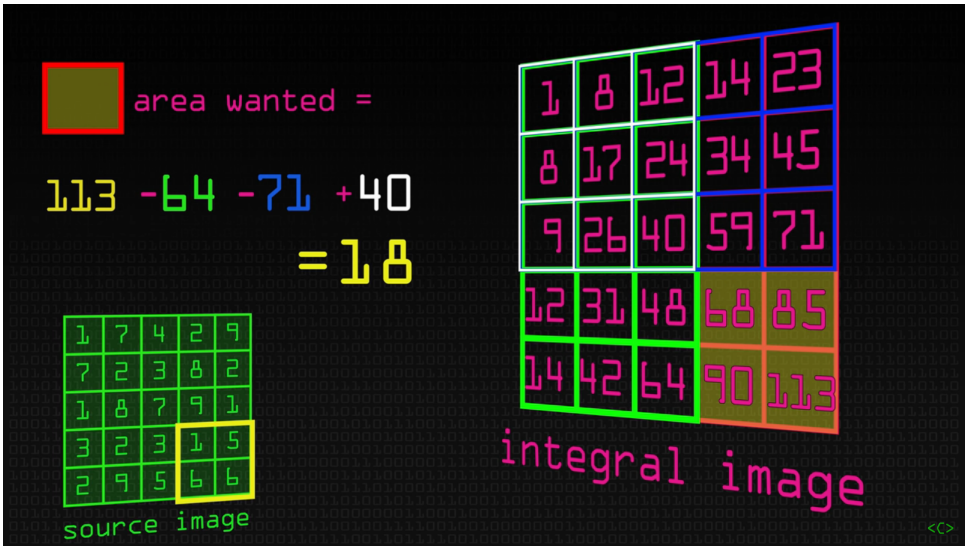

Figure 9: Finding out resultant area in an image

9 Graph and Gesture Recognition using Deep Learning

9.0.1 Deep Learning

Deep learning is a techinique of Machine learning that :

1. uses multiple layers for feature extraction and transformation. Each successive layer uses the output from the previous layer as input.
2. learns in supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis) manners.
3. learns multiple levels of representations that correspond to different levels of abstraction. The levels form a hierarchy of concepts.

Deep learning is a part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.

Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts.

Deep learning models are vaguely inspired by information processing and communication patterns in biological nervous systems yet have various differences from the structural and functional properties of human brains, which make them incompatible with neuroscience evidences.

9.0.2 Convolutional Neural Networks (CNN)

In deep learning, a convolutional neural network (CNN) is a class of deep neural networks, most commonly applied to analyze visual imagery. "CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing". They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.

Convolutional networks were inspired by biological processes. In that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field.

CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. This independence is a major advantage from prior knowledge and human effort in feature design.They have applications in image and video recognition, recommender systems, image classification, medical image analysis, and natural language processing.

1. Convolutional

   Convolutional layers apply a convolution operation to the input, passing the result to the next layer. The convolution emulates the response of an individual neuron to visual stimuli.

   Each convolutional neuron processes data only for its receptive field. Although fully connected neural networks can be used to learn features as well as classify data, it is not practical to apply this architecture to images. A very high number of neurons would be necessary, even in a shallow architecture, due to the very large input sizes associated with images, where each pixel is a relevant

variable. For instance, a fully connected layer for a small image of size 100 x 100 has 10000 weights for each neuron in the second layer. The convolution operation brings a solution to this problem as it reduces the number of free parameters, allowing the network to be deeper with fewer parameters. For instance, regardless of image size, tiling regions of size 5 x 5, each with the same shared weights, requires only 25 learnable parameters.

2. Pooling

   Convolutional networks may include local or global pooling layers, which combine the outputs of neuron clusters at one layer into a single neuron. For example, max pooling uses the maximum value from each of a cluster of neurons at the prior layer. Another example is average pooling, which uses the average value from each of a cluster of neurons from the previous layer.

3. Fully Connected

   Fully connected layers connect every neuron in one layer to every neuron in another layer. It is same as the traditional multi-layer perceptron neural network. The flattend matrix goes through a fully connected layer to classify the images

4. Receptive Fields

   In neural networks, each neuron receives input from some number of locations in the previous layer. In a fully connected layer, each neuron receives input from every element of the previous layer. In a convolutional layer, neurons receive input from only a restricted subarea of the previous layer. Typically the subarea is of a square shape (e.g. size 5 by 5). The input area of a neuron is called its receptive field. So, in a fully connected layer, the receptive field is the entire previous layer. In a convolutional layer, the receptive area is smaller than the entire previous layer.

9.0.3 Working

Let us consider the use of CNN for image classification in more detail. The main task here is acceptance of the input image and the following definition of its class. This is a skill that people learn from their birth and are able to easily determine that the image in the picture is an elephant. But the computer sees the pictures quite differently:

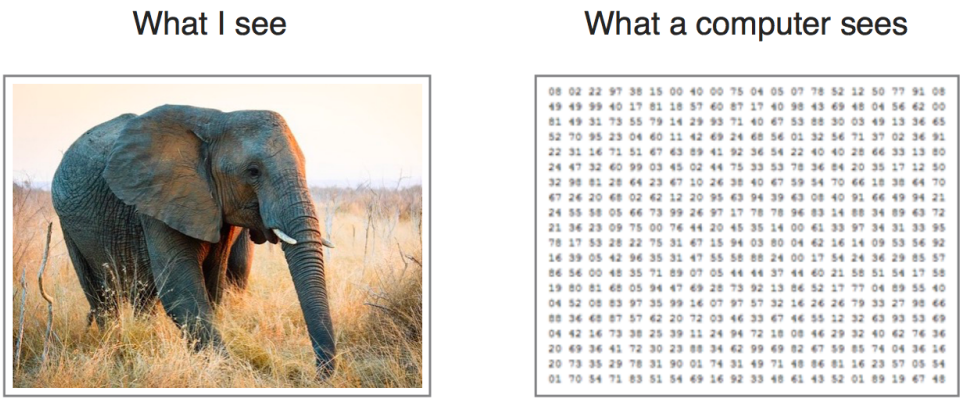

Figure 10: Finding out resultant area in an image

Instead of the image, the computer sees an array of pixels. For example, if image size is 300 x 300. In this case, the size of the array will be 300x300x3. Where 300 is width, next 300 is height and 3 is RGB channel values. The computer is assigned a value from 0 to 255 to each of these numbers. his value describes the intensity of the pixel at each point.

To solve this problem, the computer looks for the characteristics of the base level. In human understanding such characteristics are for example the trunk or large ears. For the computer, these characteristics are boundaries or curvatures. And then through the groups of convolutional layers the computer constructs more abstract concepts.

In more detail: the image is passed through a series of convolutional, nonlinear, pooling layers and fully connected layers, and then generates the output.

he image (matrix with pixel values) is entered into it. Imagine that the reading of the input matrix begins at the top left of image. Next the software selects a smaller matrix there, which is called a filter (or neuron, or core). Then the filter produces convolution, i.e. moves along the input image. The filters task is to multiply its values by the original pixel values. All these multiplications are summed up. One number is obtained in the end. Since the filter has read the image only in the upper left corner, it moves further and further right by 1 unit performing a similar operation. After passing the filter across all positions, a matrix is obtained, but smaller then a input matrix.

In order to recognize the properties of a higher level such as the trunk or large ears the whole network is needed. The network will consist of several convolutional networks mixed with nonlinear and pooling layers. When the image passes through one convolution layer, the output of the first layer becomes the input for the second layer. And this happens with every further convolutional layer.

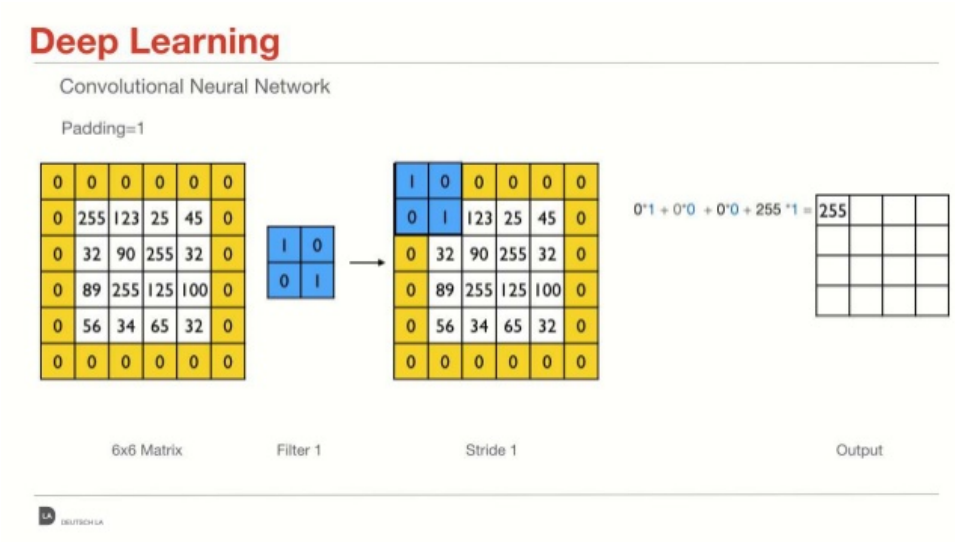

Figure 11: Finding out resultant area in an image

The nonlinear layer is added after each convolution operation. It has an activation function, which brings nonlinear property. Without this property a network would not be sufficiently intense and will not be able to model the response variable (a class label).

The pooling layer follows the nonlinear layer. It works with width and height of the image and performs a downsampling operation on them. As a result the image volume is reduced. This means that if some features have already been identified in the previous convolution operation, than a detailed image is no longer needed for further processing, and it is compressed to less detailed pictures. Attaching a fully connected layer to the end of the network results in an N dimensional vector, where N is the amount of classes from which the model selects the desired class.

10 Conclusion

This proposal introduces the structure of Next-Generation of Virtual Personal Assistants that is a new VPAs system designed to converse with a human, with a coherent structure. This VPAs system has used speech, graphics, video, gestures and other modes for communication in both the input and output channel. Also, the VPAs system will be used to increase the interaction between users and the computers by using some technologies such as gesture recognition, image/video recognition, speech recognition, and the Knowledge Base. Moreover, this system can enable a lengthy conversation with users by using the vast dialogue knowledge base. Moreover, this system can be used in different tasks such as education assistance, medical assistance, robotics and vehicles, disabilities systems, home automation, and security access control.

Also, it can be a satisfactory solution that can be used by applications, such as responding to customers, customer service agent, training or education, facilitating transactions, online shopping, travelling information, counseling, tutoring system, ticket booking, remote banking, travel reservation, Information enquiry, stock transactions, taxi bookings, and route planning etc. In the end, to achieve the final stage and all these improvements to the new system with high accuracy, we need funding from an organization that will work with us to improve the system by funding the new hardware devises that have high accuracy, as well as the tools and cloud severs that we will need for testing the new system.

11 References

1. https://medium.com/@ksusorokina/image-classification-with-convolutional-neural-networks-496815db12a8
2. https://www.youtube.com/watch?v=uEJ71VlUmMQ
3. Paul Voila and Micheal Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. 2001
4. https://www.learnopencv.com/image-recognition-and-object-detection-part1/
5. https://en.wikipedia.org/wiki/Viola-Jone_object_detection_framework