Ain Shams University.

Faculty Of Computer and Information Sciences.

Information Systems Department.

Body Language and Emotion Recognition

Prepared By:

Eman Mohamed Awad.

Shrook Ehab Attia.

Mohamed Taher Abdelsamia.

Abdallah Hossam Abdallah.

Fady Zaher Halim.

Under Supervision of:

DR. Mahmoud Mounir.

Information System Department,

Faculty of Computer and Information Sciences,

Ain Shams University.

TA. Mohamed Ashraf.

Teaching assistant,

Information System Department,

Faculty of Computer and Information Sciences,

Ain Shams University.

ACKNOWLEDGEMENT:

Before Anything we want to Thank God to achieve what we want professionally and make our road easy and send us good people to help us.

Then we want to thank our DR. Mahmoud Mounir for his appreciation, his ideas about the project, revising all our work step by step, and always was our supportive.

And we want to thank TA. Radwa Reda for her support, and always be there, helping, encouraging, supporting, searching for solutions with us, she was acting like a team member in our team.

At last, we want to thank our elder friends that helped us with tier experience in this field, give us ideas about what is new in the market, and helped us to reach our target.

ABSTRACT:

Body Language is an important element in identifying the emotional and facial expressions in human communications.

For several decades, in emotion research, body movements have been largely neglected, although they make an important contribution to emotion recognition and translation of the facial emotion.

For example, through facial expressions, body movements and speech prosody, based on the behavior of their interlocutors as well as their own communication goals, in addition to understanding the details of human interaction mechanisms.

Because of Covid-19 disease **e-learning** became more important these days, e.g., the instructor makes the lecture online (**live lecture**), so the students attend their lectures from home.

In our project we are using body language with facial expressions to identify the human action in natural scenes, e.g., a girl attending birthday party, a boy is talking with his friends in the park, …., etc.

But our main target is the instructors and their students, so in our project we are showing to the instructor the feeling of the student in the lecture from a **live website**, by **detecting** the facial expressions and body movements of every student whether he/she is (angry, fear, happy, …., etc.).

LIST OF FIGURES:

LIST OF ABBREVIATIONS:

Table of Contents

1. Introduction:

1.1. Motivation:

Body Language is more powerful than words and Emotional Expressions are important to identify the human actions, analyze their communications and reactions, so it is used to solve many problems in real life, and we use it when we speak to explain what we say.

Every day, we communicate more with our energy, body language, face, and eyes. That really is what communication is, and not many words. And it is rare that you get to explore that in a film.

In science, it has been approved that body language takes the first place in detecting the person's emotion while interacting, its name is Mehrabian's 7-38-55 Rule to Negotiate Effectively:

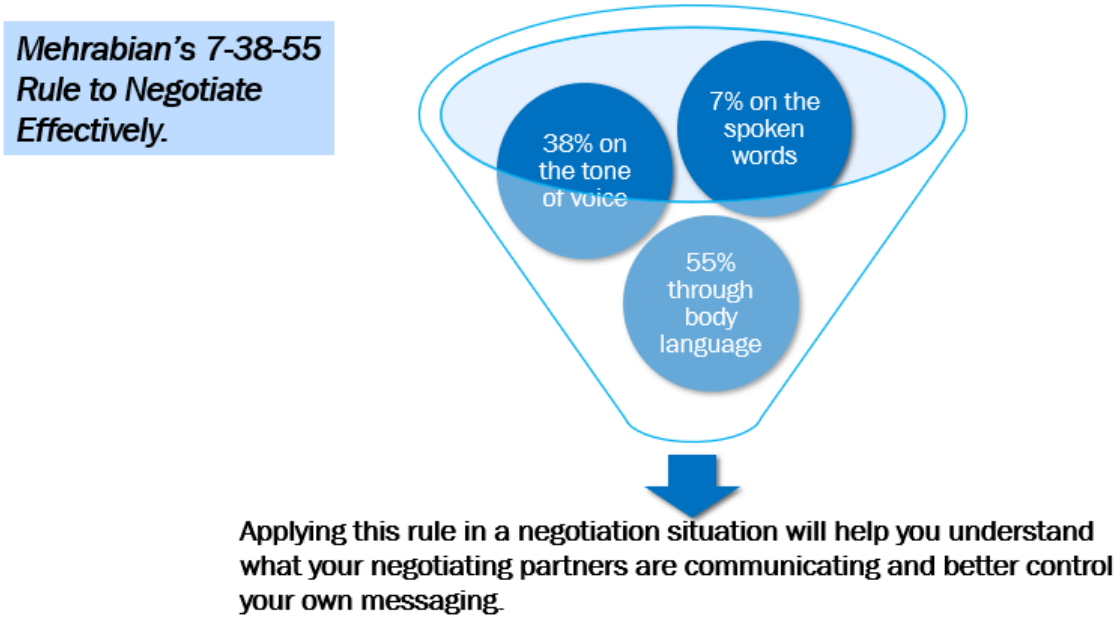


Figure 1 - Mehrabian Negotiation Rule

1.2. Problem definition:

While countries are at different points in their COVID-19 infection rates, worldwide there are currently more than 1.2 billion children in 186 countries affected by school closures due to the pandemic.

Research suggests that online learning has been shown to increase retention of information, and take less time, meaning the changes coronavirus have caused might be here to stay.

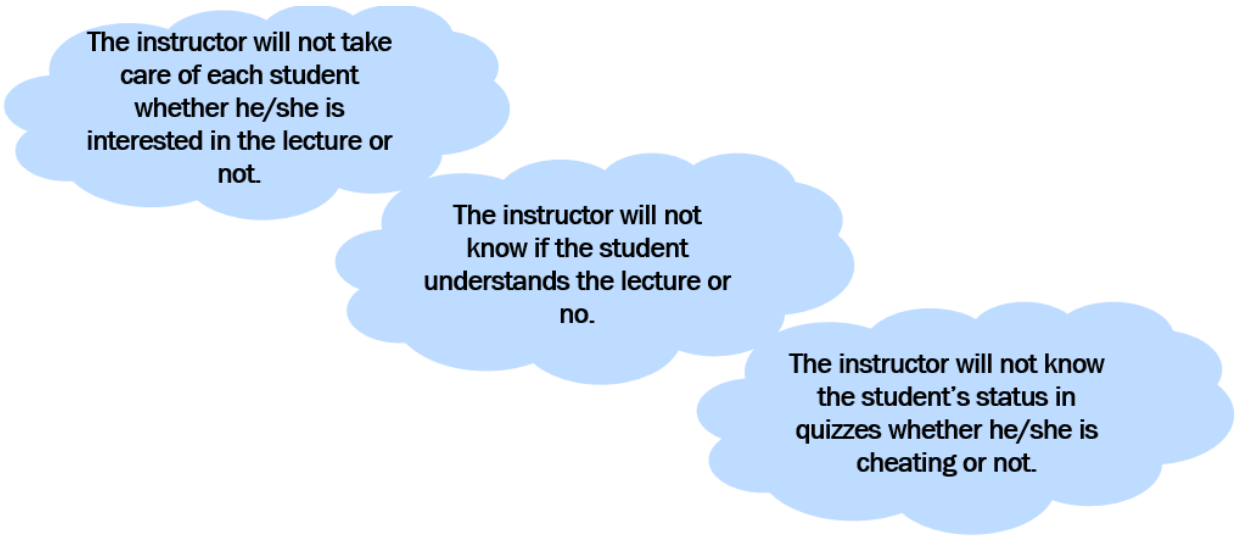


Figure 2 - Problem Definition

1.3. Objective:

we present a real time approach to emotion recognition through facial expression in live video. We employ an automatic facial feature tracker to perform face localization and feature extraction.

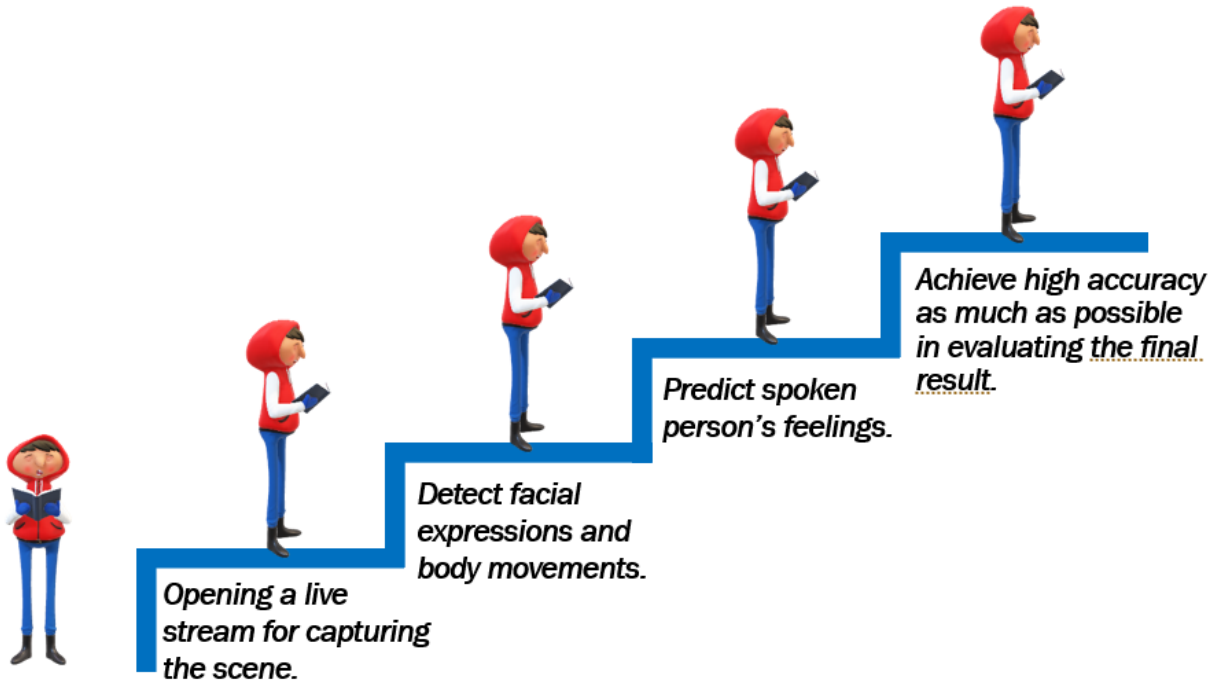


Figure 3 - System Objective

1.4. Project Benefits:

- Predict the Emotions of any human by detecting its facial expressions and body movements through live sessions.
- Instructors can use our website in their live lectures to predict their student's emotions through the lecture.
- Could be used in banks to prevent thefts by detecting their emotions when interacting with the bank clerk, also after the end of the deal with the client, we do not know whether he/she is happy or not.
- Could be used in the traffic to detect the driver's emotion if he/she is following the rules or not.
- The police may face a very important problem, which is the failure to identify the person if he/she were the perpetrator of the crime or not, so he could use this website through the investigation.
- In live interviews the HR cannot determine this person is suitable for this job or not, and whether he has confidence in his abilities or not, and this will appear through his facial expressions if he is confident in himself or is he Anxious and tense.

1.5. Document Organization:

The following Sections of this document are organized as follows:

2) Background:

- This section gives an overview of the scientific background behind the project.
- In this Section we will discuss all the techniques and approaches available for detecting human's emotions from his/her body movements and facial expressions.

3) Analysis and Design:

- This section presents the outcomes of the analysis and design phases of the project.
- It discusses the functional and nonfunctional requirements of the system and its architecture.
- It discusses the system users.

4) Implementation and Testing:

This section explains:

- The technologies and tools used in the implementation of the system.
- The role of each of them in the project.
- A simple code for the main functions.

5) User Manual:

This Section explains:

- How to operate the system and use its functionalities.
- How to use the website for detection.

6) Results:

This section shows the results of our project during the development, and a survey of peoples' prediction in some photos about the emotions in this photo

7) Conclusion and future work:

At last, this section introduces the conclusion of our work, and the future work that will be done to improve the performance of the project.

2. Background:

2.1. Description of project field:

Body language is a type of nonverbal communication in which physical behaviors, as opposed to words, are used to express, or convey the information. Such behavior includes facial expressions, body posture, gestures, eye movement, touch, and the use of space.

Although body language is an important part of communication, most of it happens without conscious awareness.

In a society, there are agreed-upon interpretations of behavior. Interpretations may vary from country to country, or culture to culture. On this note, there is controversy on whether body language is universal. Body language, a subset of nonverbal communication, complements verbal communication in social interaction.

In fact, some researchers conclude that nonverbal communication accounts for most of the information transmitted during interpersonal interactions. It helps to establish the relationship between two people and regulates interaction but can be ambiguous.

It is the relaxed facial expression that breaks out into a genuine smile – with mouth upturned and eyes wrinkled. It can be a tilt of the head that shows you are listening, sitting, or standing upright to convey interest, or directing attention with hand gestures. It can also be taking care to avoid a defensive, arms-crossed posture, or restlessly tapping your feet.

When you can "read" signs like these, you can understand the complete message of what someone is telling you. You will be more aware of people's reactions to what

you say and do. And you will be able to adjust your body language to appear more positive, engaging, and approachable.

Humans can recognize emotions through centuries, now we are making machines do the same. Emotion recognition can be done through body language, voice intonation, expressions.

Though facial emotion recognition remains the most practical method. Facial expressions are triggered for a period as a response to the internal emotional state of a person. These also display social communications.

Various attempts have been made which resulted in overcoming limitations and bringing new opportunities and to better understand and apply this simple way of human interaction in our world of computers.

There has been the usage of new technologies for capturing facial expressions, with rapid, high-resolution image acquisition, these help us to analyze and recognize in real-time the true facial emotions.

Body "language" must not be confused with sign language, as sign languages are languages and have their own complex grammar systems, as well as being able to exhibit the fundamental properties that exist in all languages. Body language, on the other hand, does not have a grammar system and must be interpreted broadly, instead of having an absolute meaning corresponding with a certain movement, so it is not a language, and is simply termed as a "language" due to popular culture.

2.2. Related Works:

Reference Name	Format	Emotions	Modalities	Samples	Accuracy
Context based emotion recognition using emotic dataset (2019).[3]	Colored images.	Discrete emotions (26), Continuous emotions (3).	Facial Expression Recognition, Body poster and gesture analysis.	(23,571) Labeled images and (34,320) annotated people.	29.5%
Real time facial emotion recognition with deep convolutional neural network (2020).[1]	Black and white images.	Discrete emotions (7).	Real-time Facial Expression Recognition.	(35,887) labeled images.	58%
Emotion Understanding in Videos Through Body, Context, and Visual-Semantic Embedding Loss (2020).[2]	Video clips.	Discrete emotions (26), Continuous emotions (3).	Body poster and gesture analysis, Context, and Visual-Semantic Embedding Loss.	(9,876) labeled video clips.	26%.

Figure 4 - Related work 1

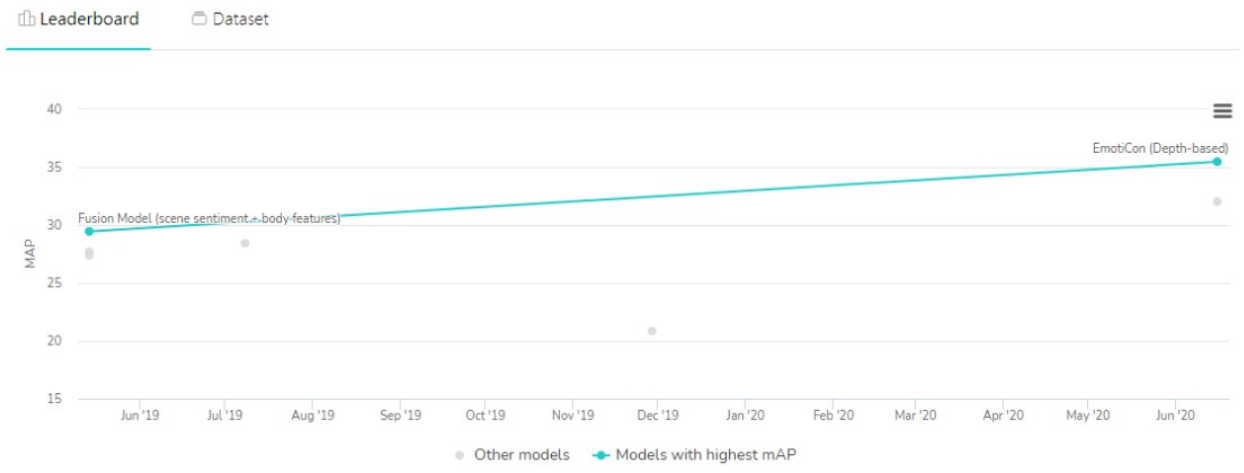


Figure 5 - Related work 2

RANK	MODEL	MAP ↑	EXTRA TRAINING DATA	PAPER	CODE	RESULT	YEAR
1	EmotiCon (Depth-based)	35.48	×	EmotiCon: Context-Aware Multimodal Emotion Recognition using Frege's Principle			2020
2	EmotiCon (GCN)	32.03	×	EmotiCon: Context-Aware Multimodal Emotion Recognition using Frege's Principle			2020
3	Fusion Model (scene sentiment + body features)	29.45	✓	Context Based Emotion Recognition using EMOTIC Dataset			2019
4	Affective Graph (GCN)	28.42	×				2019
5	Fusion Model (scene + body features)	27.70	✓	Context Based Emotion Recognition using EMOTIC Dataset			2019
6	Fusion Model	27.38	×	Context Based Emotion Recognition using EMOTIC Dataset			2019
7	CAER-Net (Adaptive Fusion)	20.84	×	Context-Aware Emotion Recognition Networks			2019

Figure 6 - Related work 3

3. Analysis and Design:

3.1. System overview:

3.1.1. System architecture:

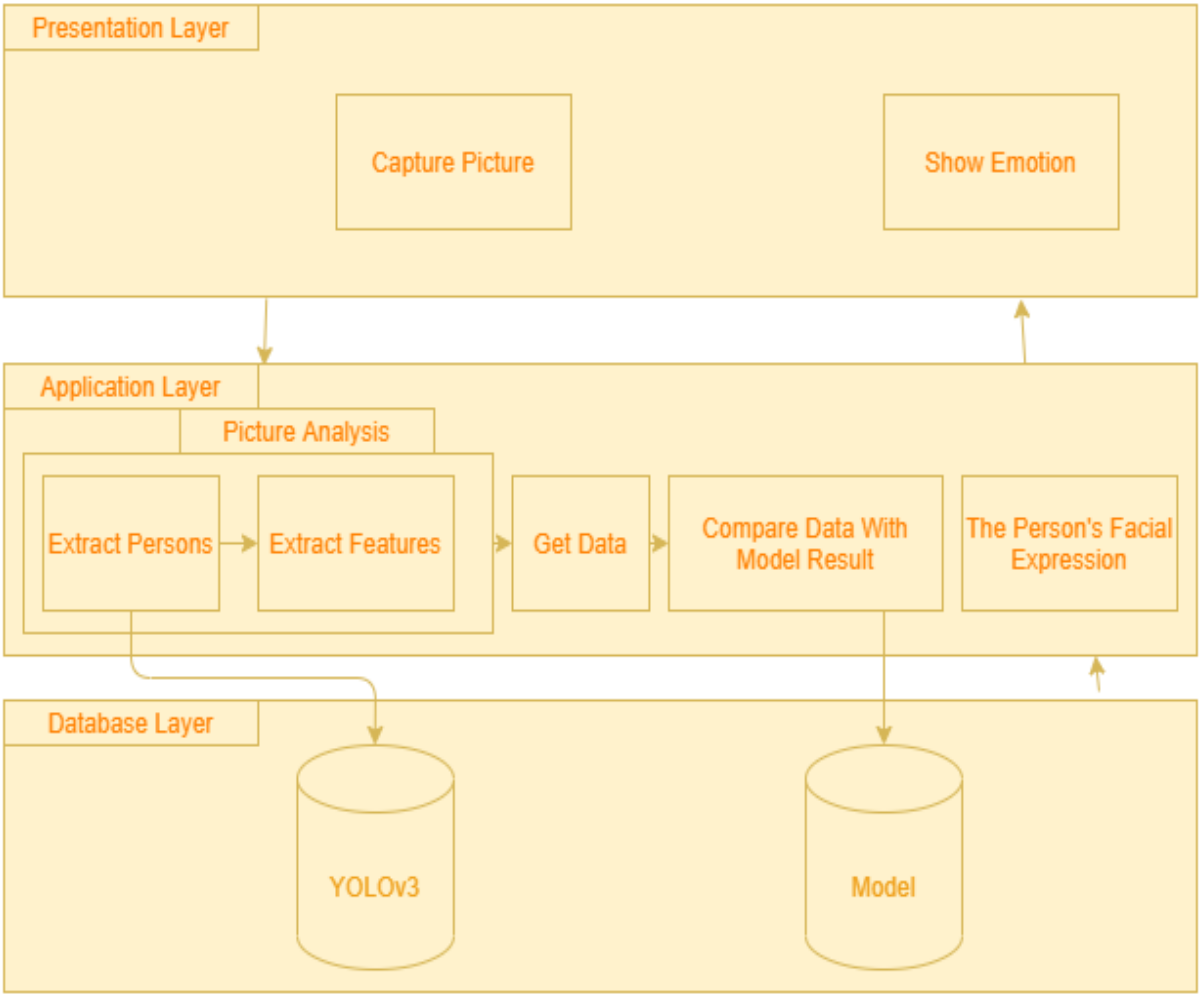


Figure 7 - System Architecture

1. **Presentation Layer:**
    - Picture Link: User Input (Picture Live/Video).
    - Show Result: System Output (Emotions).
    - Internet: To take Picture Live need to connect to the internet.

2. **Application Layer:**
    - Preprocessing: prepare data for Picture Analysis.
    - Persons Extract: extract all person from picture.
    - Features Extract: extract all features from all person enter the picture.
    - Model Classification: classify persons that enter picture to Emotions or classes that suitable for each person.

3. **Database Layer:**
    - Yolov3: detect object entre picture.
    - Model: Show Emotions of Person.

3.1.2. System Time Plan:

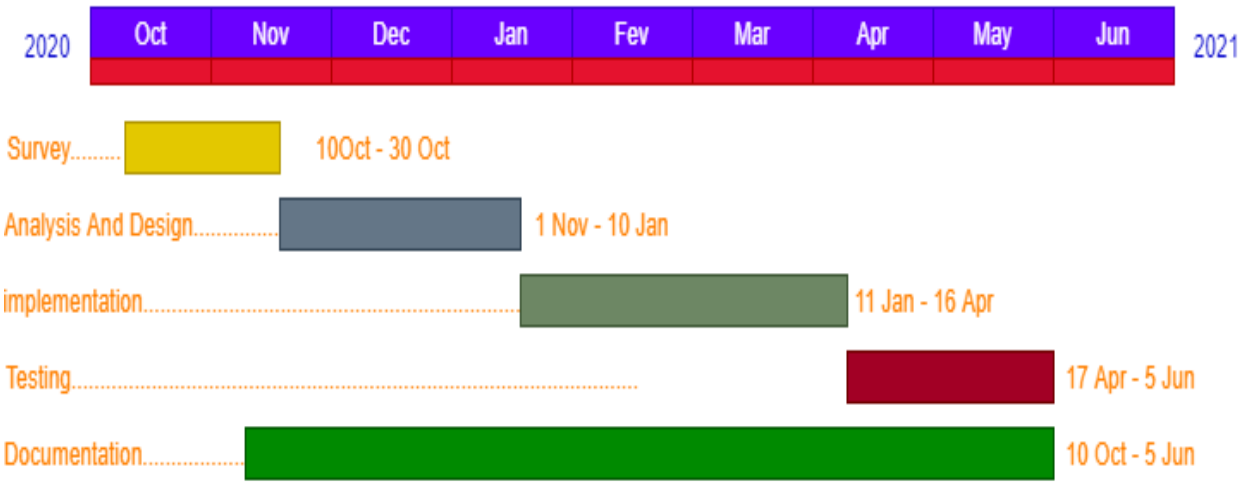


Figure 8 - System Time Plan

3.1.3. Functional Requirements:

1) Capture Image Method:

This function takes a picture from live then send to get box Method then infer method and finally return emotion of persons.

2) Upload Video Method:

This function takes a Video and divide each frame into 30 pictures then send to get box Method then infer method and finally return emotion of persons.

3) Get BBox of each person:

This function takes a picture and puts a box on each person inside this picture.

4) inference method:

This function takes a picture Which contains one person, then you compare it with the model and take the result, write it on the person and return the image.

5) Get Features from Dataset:

This function runs the model on the data, then returns the feature map, and then converts the feature to a file.

6) Saving the model weights:

This function saves model weights and check if number of classes suitable for weights.

7) Prepare Emotic model:

Calculate features of face in object and features of body in another object.

8) Call optimizer:

Use Adam optimizer to update weights to improve accuracy.

9) Train and validate the model:

This is the Method to train the model on the data
Firstly, the training data and their losses are calculated, then the results are recorded, then validate the data and calculate their losses.

10) Test the model:

This function takes the feature of the person then
Start making a test for the categories of this features, it calculates the probability ratio that these features belong to these categories, Until the percentage is higher, the result is stored, and the classes of this person are returned.

3.1.4. Non-Functional Requirements:

1) Usability:

This focuses on the appearance of the user interface and how people interact with it. What color are the screens and how big are the buttons.

2) Reliability / Availability:

This focuses on the uptime requirements.

3) Scalability:

As needs grow, can the system handle it? For physical installations, this includes spare hardware or space to install it in the future.

4) Performance:

This focuses on fast that need to operate.

5) Supportability:

Is support provided in-house or is remote accessibility for external resources required?

6) Security:

What are the security requirements, both for the physical installation and from a cyber perspective?

3.2. System Users:

3.2.1. Intended Users:

- The System is Built Specifically for e-learning Instructors and their Students.
- Each Student should allow permission for the website to record video and audio.
- Each Instructor can capture the student photo at the time he/she wants or can upload a video of the students.

3.2.2. User Characteristics:

The User should know how to use computers and deal with websites at least.

3.3. System Analysis and Design:

3.3.1. Use Case Diagram:

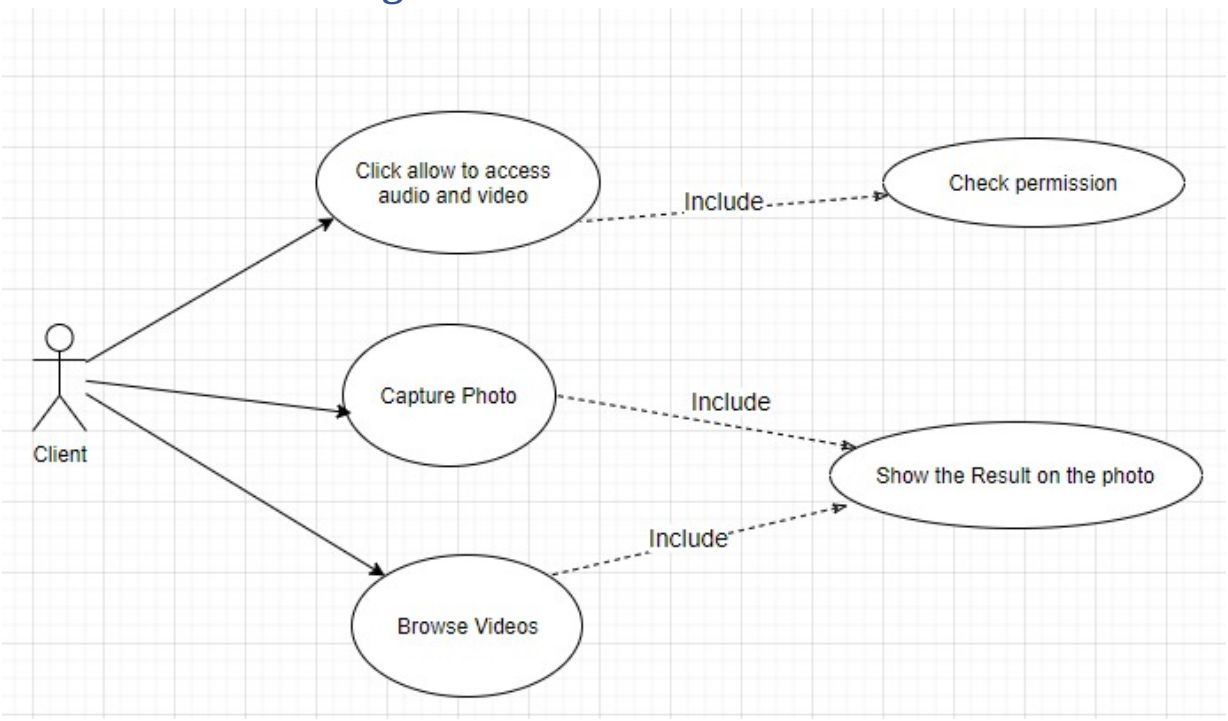


Figure 9 - System Use Case Diagram

- **Description:**
  1) Allow accessing audio and video: The user clicks allow to the system to access its audio and video to open the live camera.
  2) Capture Photo: The user clicks at capture Photo to take a snapshot from the live camera then sending it to the model API to make predictions on it.
  3) Browse video: The user clicks to upload video from his/her device to pass it and make predictions for each frame.
  4) Check permission: the system check if the access of video and audio are enabled or not.
  5) Show the Result on the Photo: user can see the Photo after making prediction on photo or each frame of the video.

3.3.2. Class Diagram:

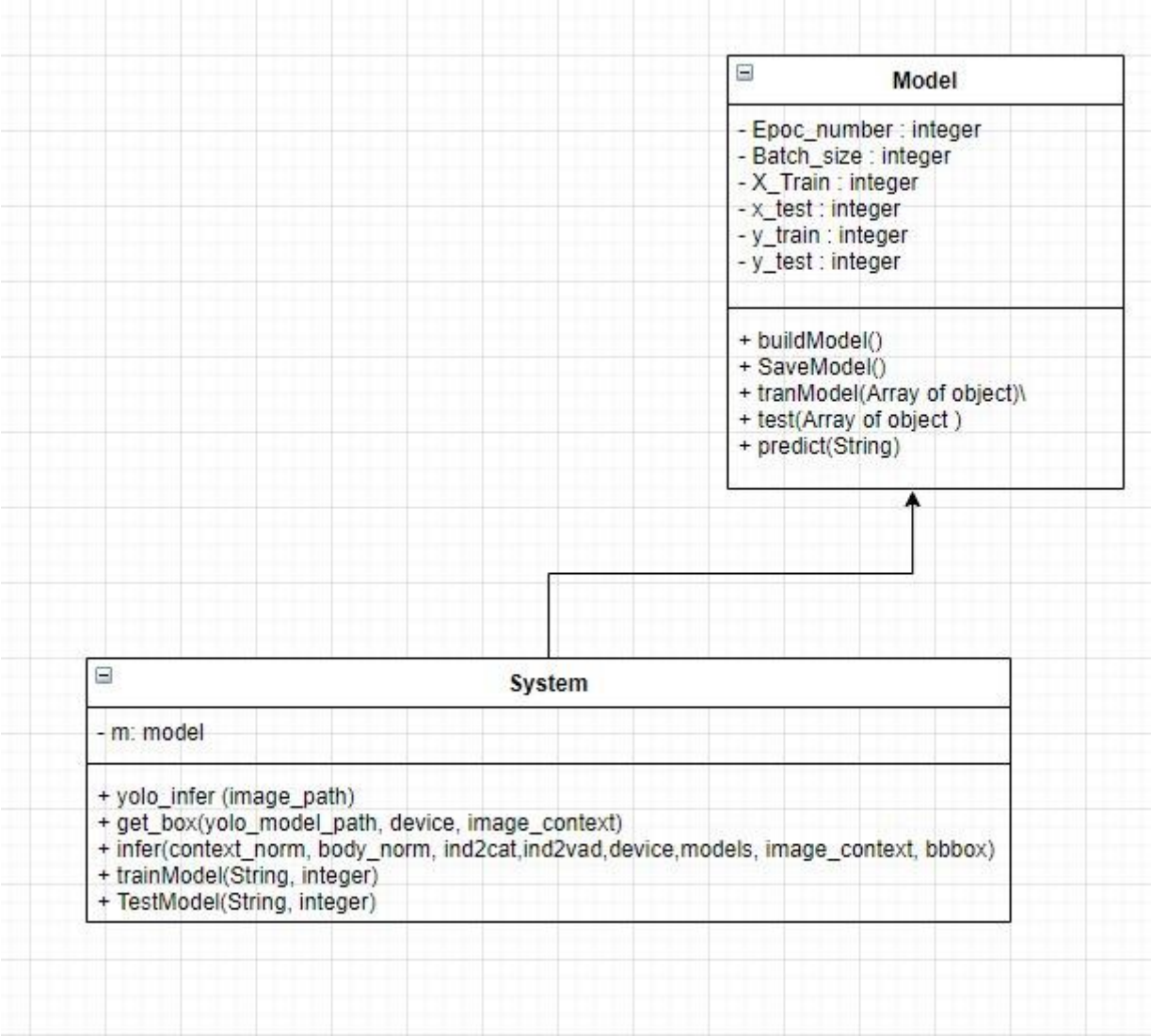


Figure 10 - System Class Diagram

3.3.3. Sequence Diagrams:

- **Sequence Diagram of the whole system:**

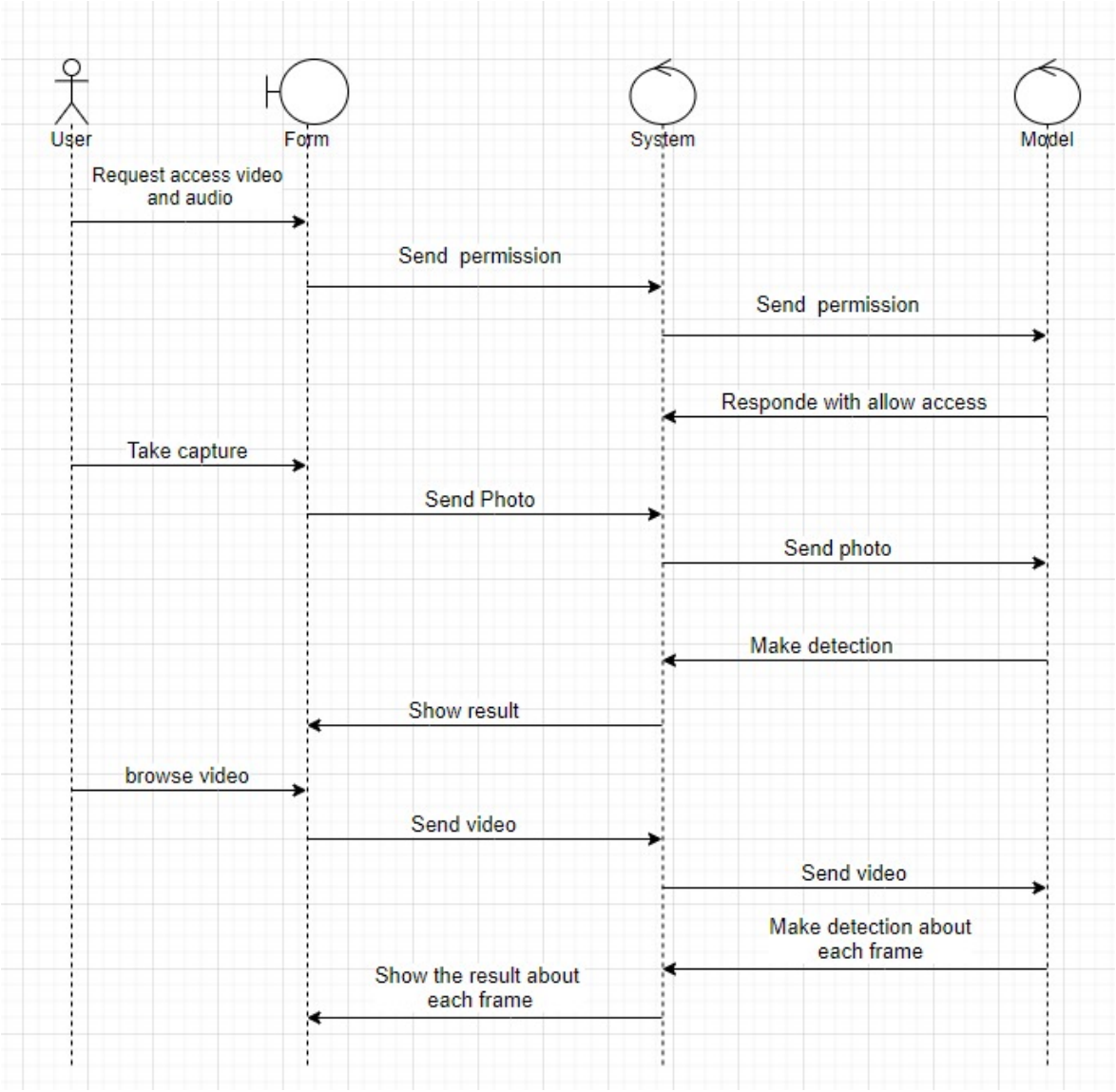


Figure 11 - System Sequence Diagram

- **Sequence Diagram for accessing audio and video:**

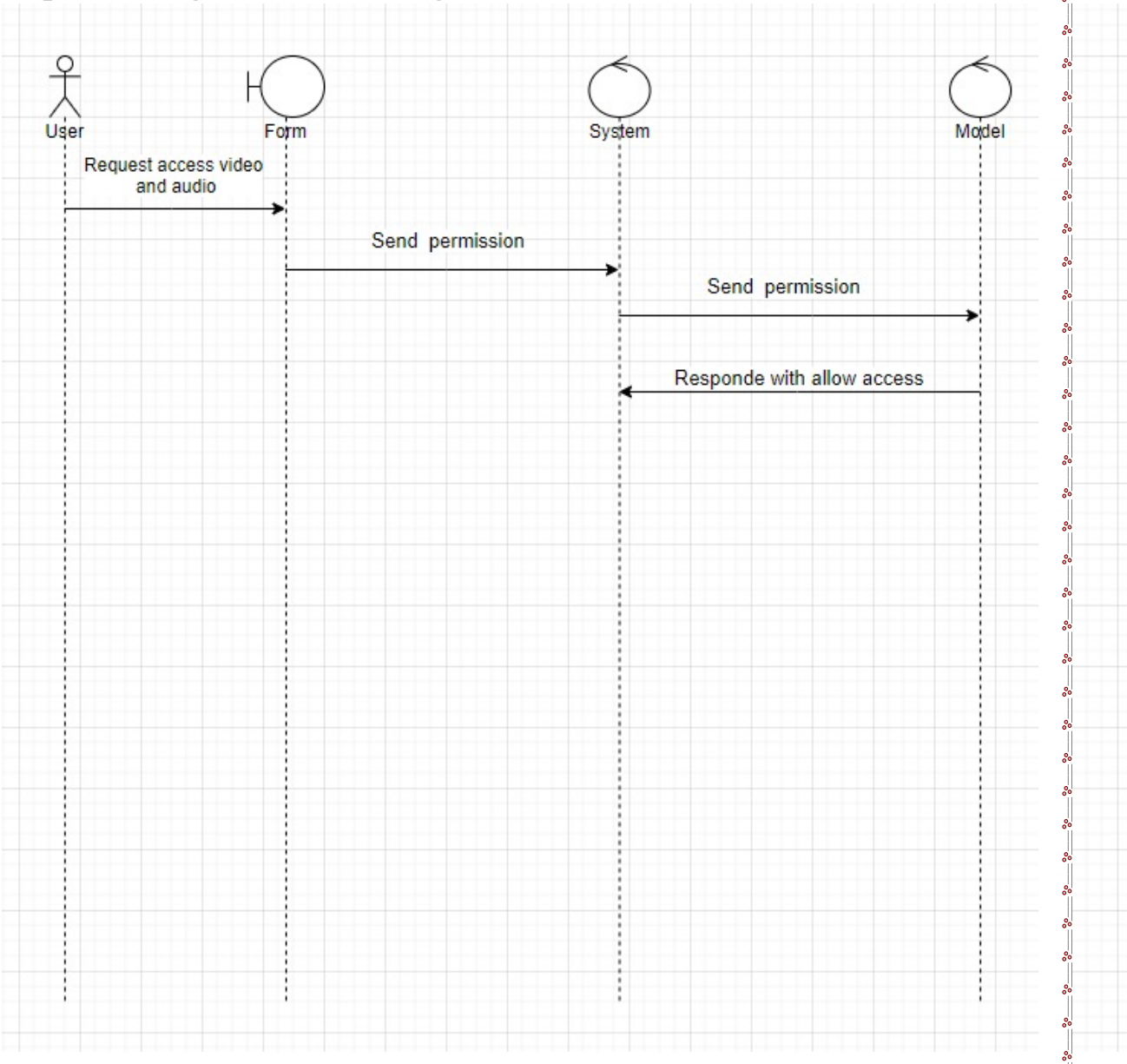


Figure 12 - Sequence Diagram for accessing audio and video

- **Sequence Diagram for Capturing photo from live camera:**

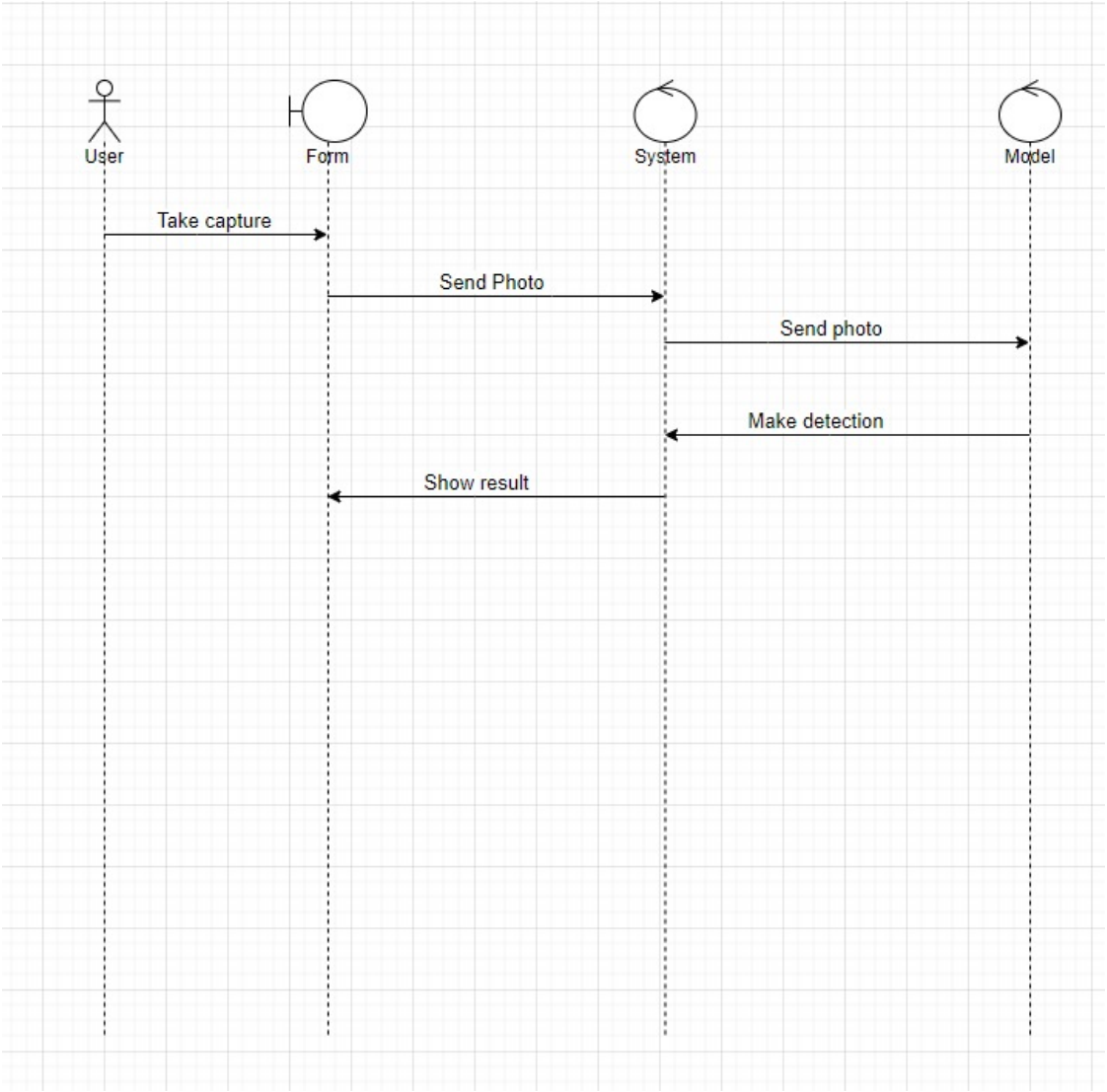


Figure 13 - Sequence Diagram for Capturing photo from live camera

- **Sequence Diagram for uploading photo or video from local pc:**

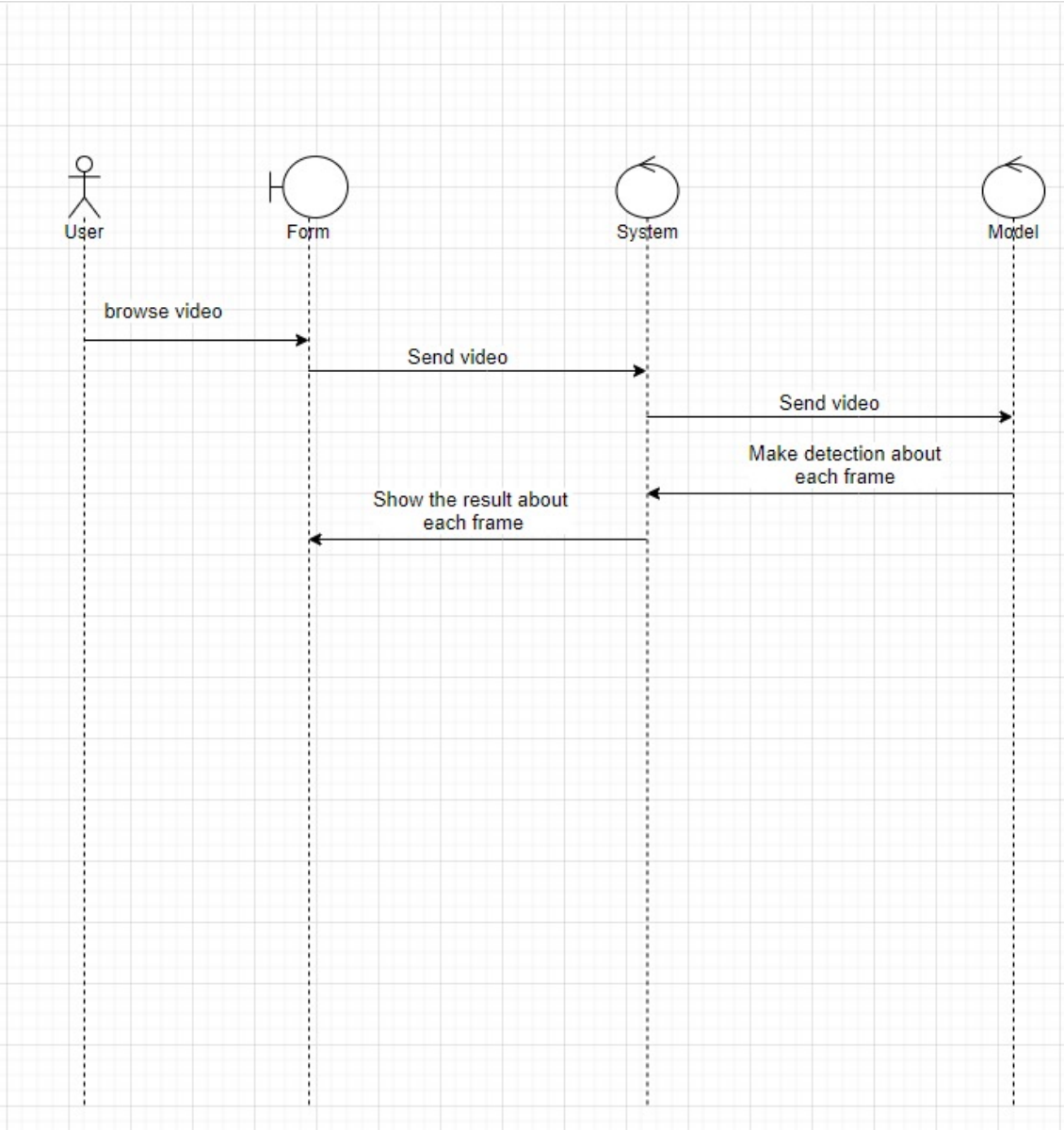


Figure 14 - Sequence Diagram for uploading photo or video from local pc

4. Implementation and Testing:

4.1. Development Environment:

1) **PyCharm Community:**

PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains (formerly known as IntelliJ). It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as data science with Anaconda.

2) **Visual Studio Code:**

Visual Studio Code is a source-code editor made by Microsoft for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git. Users can change the theme, keyboard shortcuts, preferences, and install extensions that add additional functionality.

3) **Colab:**

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether you're a student, a data scientist or an AI researcher, Colab can make your work easier.

4) Programming Languages:

- Python : is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

  - Why Python not MATLAB?

    Many data science companies prefer to use python rather than MATLAB because:

    1) It is free:

       This one is simple, but one of the most important ones for anyone working outside academia. As MATLAB is an expensive software, many companies only have one (if any) license. For this reason, using a free alternative might be attractive for many companies.

    2) It is open source:

       Having a language where everyone can contribute to the development of it means new features will constantly be added and bugs will be fixed. It also means you can go and inspect exactly how the functions you use works. This also aids in the growth of the language and helps making a big community of advanced users and contributors.

    3) It is the future:

       The popularity of Python has just kept increasing in an exponential fashion over the past years. Having a more popular language means it is easier to find answers to any

questions you may have, and to find code examples of what you need.

4) It has more features:

   Unlike MATLAB, Python is not just a scripting language for math, it is also an imperative and function language which can be used for crawling webservers, controlling external devices or making user interfaces.

5) It is portable:

   Python is, just like MATLAB, a cross-platform, language which can run on al OSs, even embedded systems having a small Linux kernel. Deploying Python code is also easier, you just need to install python (which comes by default in many OSs) and not deal with having the right version of MATLAB runtime. This also means it is super easy to deploy Python code to servers.

6) It is THE go-to language for machine learning:

   With the increasing popularity of machine learning and AI, Python is light years ahead of MATLAB, as all major frameworks are based on Python: TensorFlow, Keras, PyTorch, Scikit-learn. And since all AI research is made using these frameworks it is way easier to find state of the art algorithms for Python than MATLAB.

7) It is highly flexible:

   In Python, there are many ways to achieve the same functionality. Some are of course more efficient than others but having a language which allows you to do things the way that suits you is highly appreciated.

8) It allows for using different IDEs:

When using MATLAB, you are forced to use the MATLAB IDE shown in the image below. Luckily, the MATLAB IDE works quite well, but you are also quite limited to the features that MATLAB has chosen to implement: For instance, is Git support quite poor. As Python can be compiled from the command-line many different IDEs are available, from simple text editors to full-fledged MATLAB-like solutions.

9) Simpler (prettier) language:

Even though it might not appear so at first sight, Python can produce much simpler, and thus prettier, code than MATLAB. One such example is in for-loops where you can get both the index I and its item when iterating over an array.

10) Named arguments:

Named arguments in Python lets you call a function ala. mean (X, axis=1), where you in MATLAB would write mean(X,1) i.e., it would be unclear what the 1 is used for. This makes the code much easier to read and debug.

- Django : is a Python-based free and open-source web framework that follows the model–template–views (MTV) architectural pattern. It is maintained by the Djangooftware Foundation (DSF), an American independent organization established as a 501(c)(3) non-profit.

    - Why Django?

      Lots of businesses prefer sites built with Django because they are:

      1) Super-fast. Django helps you turn ideas into products in the shortest possible time thanks to its simple syntax.
      2) Fully loaded. There are dozens of extras and packages, so you can carry out all kinds of common tasks from user authentication and authorization to content administration.
      3) Versatile. You can use Django for almost any project, from CMSs to e-commerce apps to on-demand delivery platforms.
      4) Secure. With Django, you can prevent common security issues including cross-site request forgery, cross-site scripting, SQL injection, and clickjacking.
      5) Scalable. Django lets you can scale your website fast so it can meet high traffic demands.

- HTML5 : is a markup language used for structuring and presenting content on the World Wide Web. It is the fifth and last major HTML version that is a World Wide Web Consortium (W3C) recommendation. The current specification is known as the HTML Living Standard. It is maintained by a consortium of the major browser vendors (Apple, Google, Mozilla, and Microsoft), the Web Hypertext Application Technology Working Group (WHATWG).

- CSS3 : is the latest evolution of the Cascading Style Sheets language a aims at extending CSS2.1. It brings a lot of new features and additions, like rounded corners, shadows, gradients, transitions, or animations, as well as new layouts like multi-columns, flexible box or grid layouts.

- Bootstrap : is a free and open-source CSS framework directed at responsive, mobile-first front-end web development. It contains CSS- and (optionally) JavaScript-based design templates for typography, forms, buttons, navigation, and other interface components.

- JavaScript often abbreviated as JS : is a programming language that cforms to the ECMAScript specification. JavaScript is high-level, often just-in-time compiled, and multi-paradigm. It has curly-bracket syntax, dynamic typing, prototype-based object-orientation, and first-class functions.

- jQuery : is a JavaScript library desied to simplify HTML DOM tree traversal and manipulion, as well as event handling, CSS animation, and Ajax. It is free, open-source software using the permissive MIT License. As of May 2019, jQuery is used by 73% of the 10 million most popular websites. Web analysis indicates that it is the most widely deployed JavaScript library by a large margin, having at least 3 to 4 times more usage than any other JavaScript library.

5) Frameworks:

- Django REST Framework: is a powerful and flexible toolkit for building Web APIs which can be used to Machine Learning model deployment. With the help of Django REST framework, cplex machine learning models can be easily used just by calling an API endpoint.
- OpenCV: (Open-Source Computer Vision Library) is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then Itseez (which was later acquired by Intel). The library is cross-platform and free for use under the open-source Apache 2 License. Starting with 2011, OpenCV features GPU acceleration for real-time operations.
- TensorFlow: is a free and open-source software library for machine learning. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks. TensorFlow is a symbolic math library based on dataflow and differentiable programming. It is used for both research and production at Google.

- PyTorch: is an open-source machine learning library based on the Torch library, used for applications such as computer vion and natural language processing, primarily developed by Facebook's AI Research lab (FAIR). It is free and open-source software released under the Modified BSD license. Although the Python interface is more polished and the primary focus of development.
- Scikit-learn: (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient booing, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- JSON : is an open standard file format and data interchange format that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and arrays (or other serializable values). It is a very common data format, with a diverse range of applications, one example being web applications that communicate with a server.

4.2. Detailed Description for Main Functions:

4.2.1. Model Functions:

1) Load data from google drive:

- Load Data:

  Load the data from google drive.

- Convert model weights:

  If the model extracted new features from a new input image it will update the old weights with the new extracted features weights.

- Save model weights:

  Save the new extracted features weights.

- Load preprocessed data:

  Load the data from older versions that previously published on google drive if colab failed.

2) Calculate error percentage in code:

- Loss functions:

  There are 3 types of Loss Functions:

  1. DiscreteLoss:

     Calculate the loss of the discrete emotions (26 Emotions) weights that have been calculated previously by using this equation:

     $$Loss = (Xpred - Xtarget)^2 * Weights$$

2. ContinuousLoss_SL1:

   Calculate the loss of the Continuous emotions (VAD) weights that have been calculated previously by using this equation:

   $$Loss = |(Xpred - Xtarget)^2 * Weights|$$

3. ContinuousLoss_L2:

   Another Function to Calculate the loss of the Continuous emotions (VAD) weights that have been calculated previously by using this equation:

   $$Loss = |0.5 * (Xpred - Xtarget)^2 * Weights|$$

- Prepare optimizer:

  After calculating the weights, the model can execute the output with bad accuracy so the call the optimizer to improve the weights and the model accuracy.

3) Emotions Prediction:

The dataset contains 23,571 images, and 34,320 annotated people. And is divided into 3 divided the images into three sets: Training (70%), Validation (10%), and Testing (20%)

- Model training and validation:

  This function to train the model on the images in the dataset and validate from the features that extracted.

The model begins with loss equal to Zero then begin extracting the features from the images.

- Related functions:

1. DiscreteLoss:

   To calculate the Discrete loss of the feature's weights that have been extracted from training and validation phases.

2. ContinuousLoss_SL1:

   To calculate the Continuous loss of the feature's weights that have been extracted from training and validation phases.

3. ContinuousLoss_L2:

   To calculate the Continuous loss of the feature's weights that have been extracted from training and validation phases.

4. Prepare optimizer:

   Call the optimizer to improve the accuracy of the model.

- Model testing:

  This function to test the model on the images in the dataset to identify how much the model is trained well and calculate the MAP (the Accuracy of the model).

- Get BBox of persons from image:

  This function that detects the bounding box of each person in the image to send it coordination (X, Y, Width, Height) to the model to extract the features form it.

- Predict emotions:

  After extracting the features and each person from the image, this function takes a new image and send it to the related functions and after getting the output begin to compare the extracted features with the features of the model that has been trained on.

  - Related Functions:

    1. Get BBox of persons from image:

       This function to get the persons from the image then output its coordination.

4.2.2. Deploy the Model:

When Django project is created it is run on local server, so we publish our model on it using Django REST Framework.

4.2.3. Website Functions:

- **Allow accessing audio and video:**

  The user clicks allow to the system to access its audio and video to open the live camera.

- **Capture Photo:**

  The user clicks at capture Photo to take a snapshot from the live camera then sending it to the model API to make predictions on it.

- **Browse video:**

  The user clicks to upload video from his/her device to pass it and make predictions for each frame.

- **Check permission:**

  The system check if the access of video and audio are enabled or not.

- **Show the Result on the Photo:**

  The user can see the Photo after making prediction on photo or each frame of the video.

4.3. Techniques and Algorithms:

4.3.1. YOLOv3 Algorithm:

- **What IS YOLOv3 Model:**

  is a real-time object detection algorithm that identifies specific objects in videos, live feeds, or images. YOLO is implemented using the Keras or OpenCV deep learning libraries.

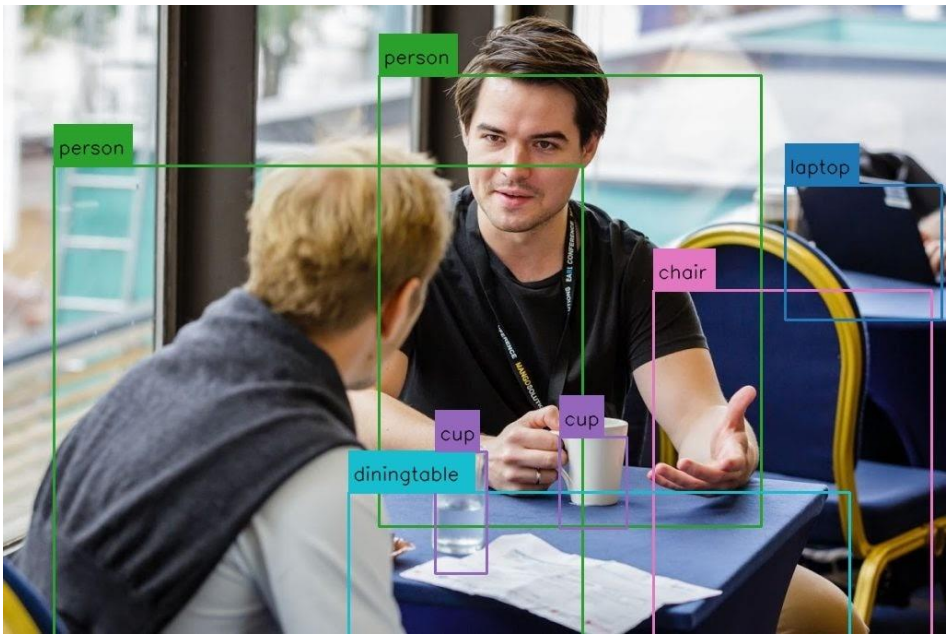


Figure 15 - YOLOv3 Object Detection Algorithm

- **Model Architecture:**

YOLO is a Convolutional Neural Network (CNN) for doing object detection. CNNs are classifier-based systems that can process input images as structured arrays of data and identify patterns between them. YOLO has the advantage of being much faster than other networks and still maintains accuracy.

It allows the model to look at the whole image at test time, so its predictions are informed by the global context in the image. YOLO and other convolutional neural network algorithms “score” regions based on their similarities to predefined classes. High-scoring regions are noted as positive detections of whatever class they most closely identify with. For example, in a live feed of traffic, YOLO can be used to detect different kinds of vehicles depending on which regions of the video score highly in comparison to predefined classes of vehicles.

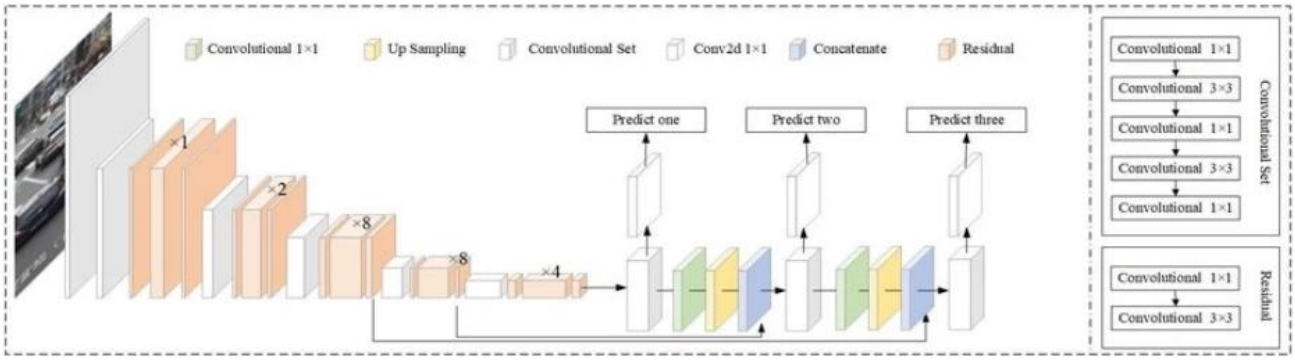
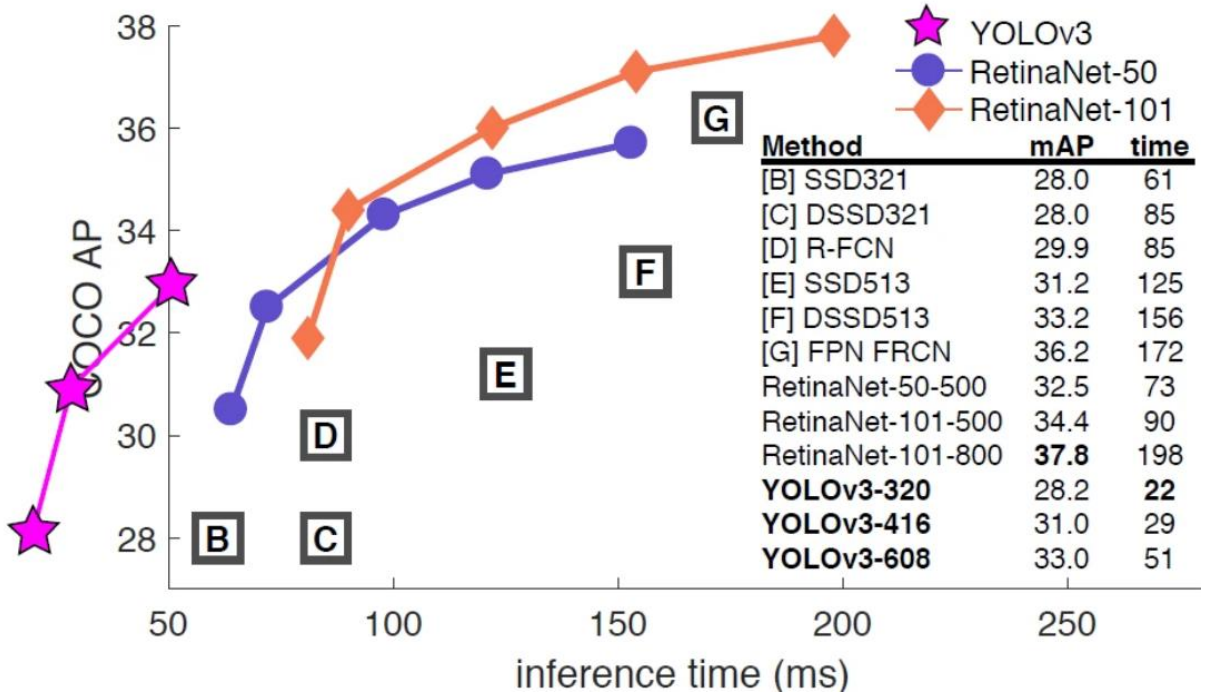


Figure 16 - YOLOv3 Object Detection Model Architecture

- **Model Speed:**

YOLOv3 is fast and accurate in terms of mean average precision (mAP) and intersection over union (IOU) values as well. It runs significantly faster than other detection methods with comparable performance (hence the name – You only look once). Moreover, you can easily trade-off between speed and accuracy simply by changing the size of the model, and no retraining required.



Method	mAP	time
[B] SSD321	28.0	61
[C] DSSD321	28.0	85
[D] R-FCN	29.9	85
[E] SSD513	31.2	125
[F] DSSD513	33.2	156
[G] FPN FRCN	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	**37.8**	198
YOLOv3-320	28.2	**22**
YOLOv3-416	31.0	29
YOLOv3-608	33.0	51

Figure 17 - YOLOv3 Object Detection Model Speed

4.3.2. ResNet-50 Algorithm:

- **Model Architecture:**

  is an improved convolutional neural network that is 50 layers deep. You can load a pretrained version of the network trained on more than a million images, make convolution function, and make pooling to extract the important feature and after that make training.

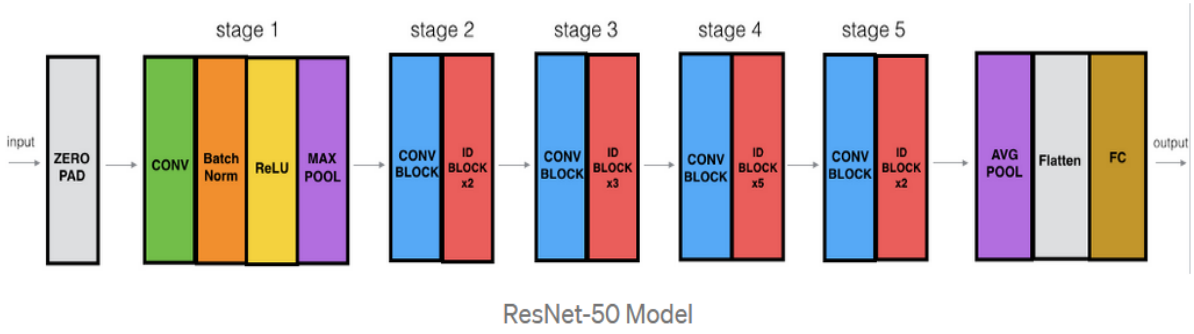


Figure 18 - ResNet-50 Algorithm Model Architecture

- **The Strength of ResNet (Skip Connection)**

  ResNet first introduced the concept of skip connection. The diagram below illustrates skip connection. The figure on the left is stacking convolution layers together one after the other. On the right we still stack convolution layers as before, but we now also add the original input to the output of the convolution block. This is called skip connection.

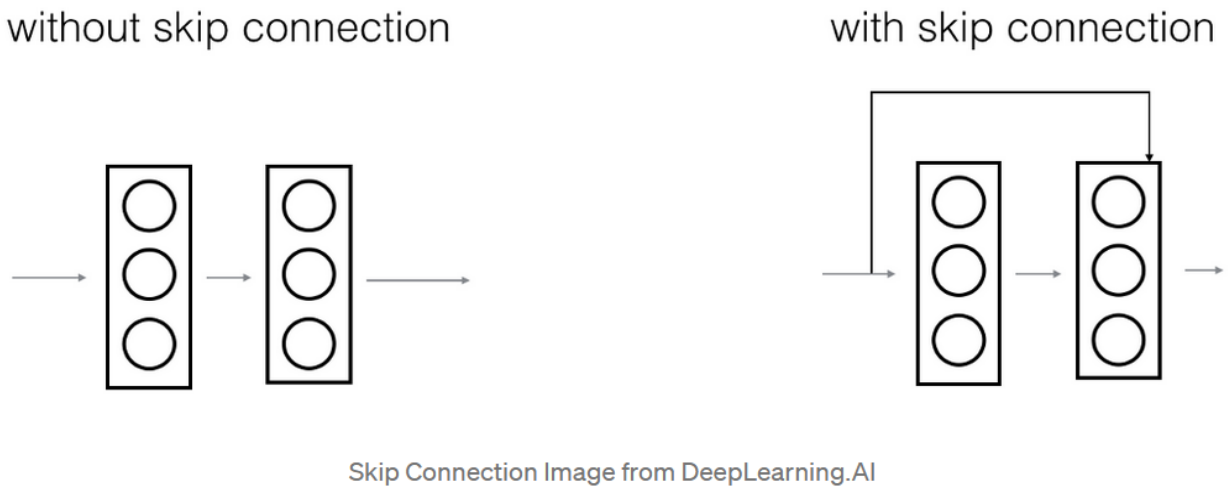


Figure 19 - The Strength of ResNet-50 Algorithm

- **Reason of using skip connection:**

  1) They mitigate the problem of vanishing gradient by allowing this alternate shortcut path for gradient to flow through.
  2) They allow the model to learn an identity function which ensures that the higher layer will perform at least as good as the lower layer, and not worse.

  In fact, since ResNet skip connections are used in a lot more model architectures like the Fully Convolutional Network (FCN) and U-Net. They are used to flow information from earlier layers in the model to later layers. In these architectures they are used to pass information from the down sampling layers to the up-sampling layers.

4.3.3. Adam Optimizer Technique:

- **What is Adam?**

  is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure (an optimization algorithm used to train machine learning models by minimizing errors between predicted and actual results) to update network weights iterative based in training data.

- Adam is a popular algorithm in the field of deep learning because it achieves good results fast.
- Empirical results demonstrate that Adam works well in practice and compares favorably to other stochastic optimization methods.

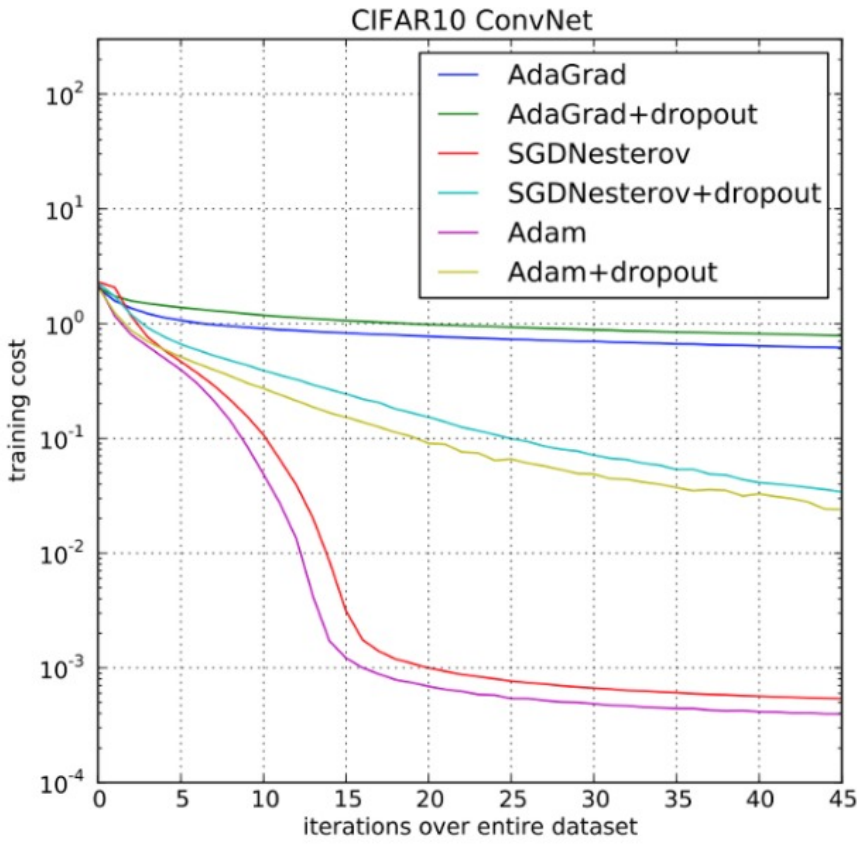


Figure 20 - Comparison between Adam Optimizer and other Optimizers

- In the original paper, Adam was demonstrated empirically to show that convergence meets the expectations of the theoretical analysis. Adam was applied to the logistic regression algorithm on the MNIST digit recognition and IMDB sentiment analysis datasets, a Multilayer Perceptron algorithm on the MNIST dataset and Convolutional Neural Networks on the CIFAR-10 image recognition dataset. They conclude:
    - Using large models and datasets, we demonstrate Adam can efficiently solve practical deep learning problems.

5. User Manual:

Step (1): Allow accessing audio and Video:

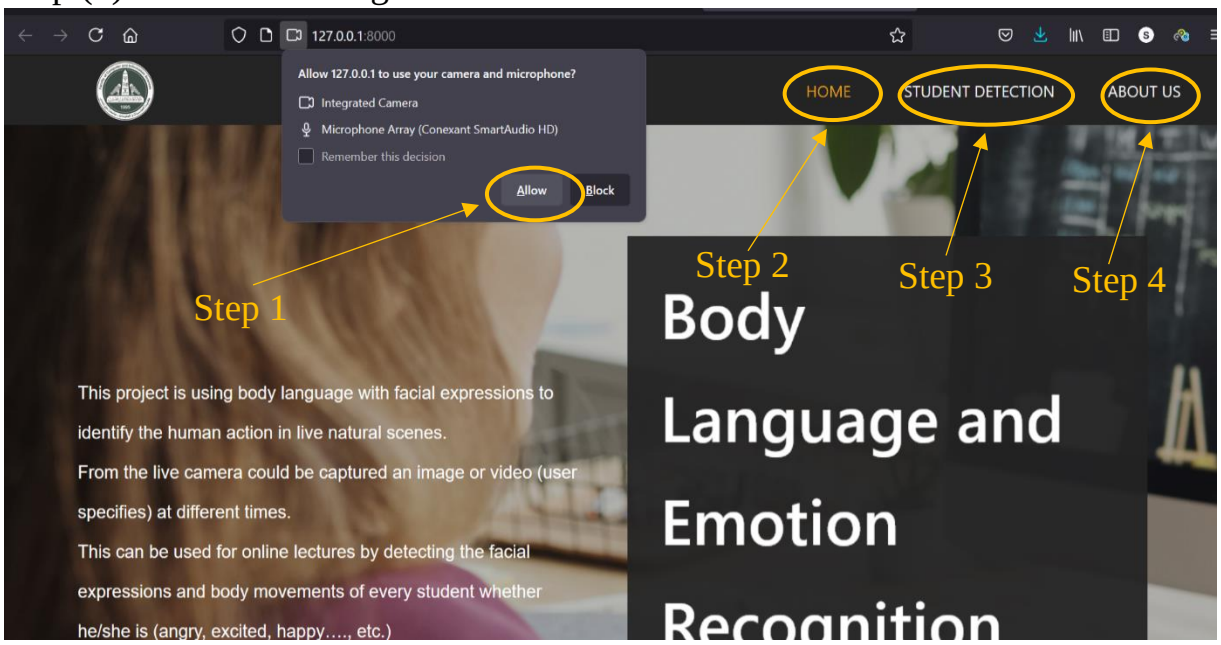


Figure 21 - User Manual Allow accessing audio and Video:

Step (2): Home Section:

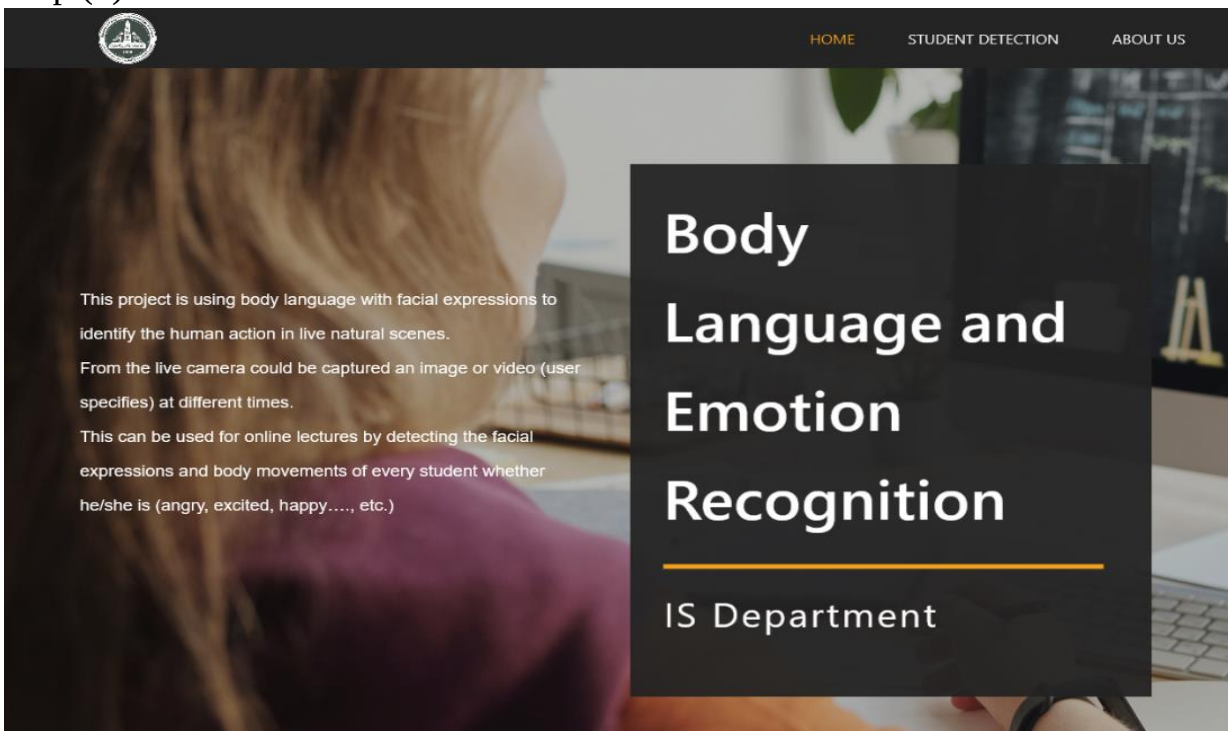


Figure 22 - User Manual Home Section

Step (3): Student Detection Section:

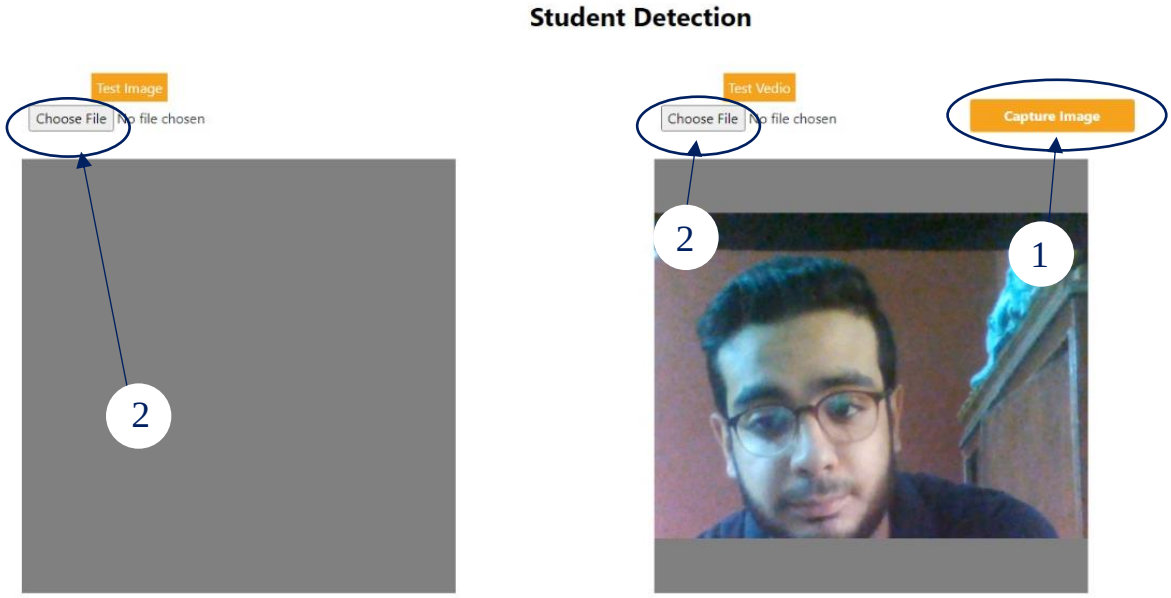


Figure 23 - User Manual Student Detection Section

1. Capture image button

to capture an image of the user from live camera and send it to the model to get results then show the user image with its emotions.

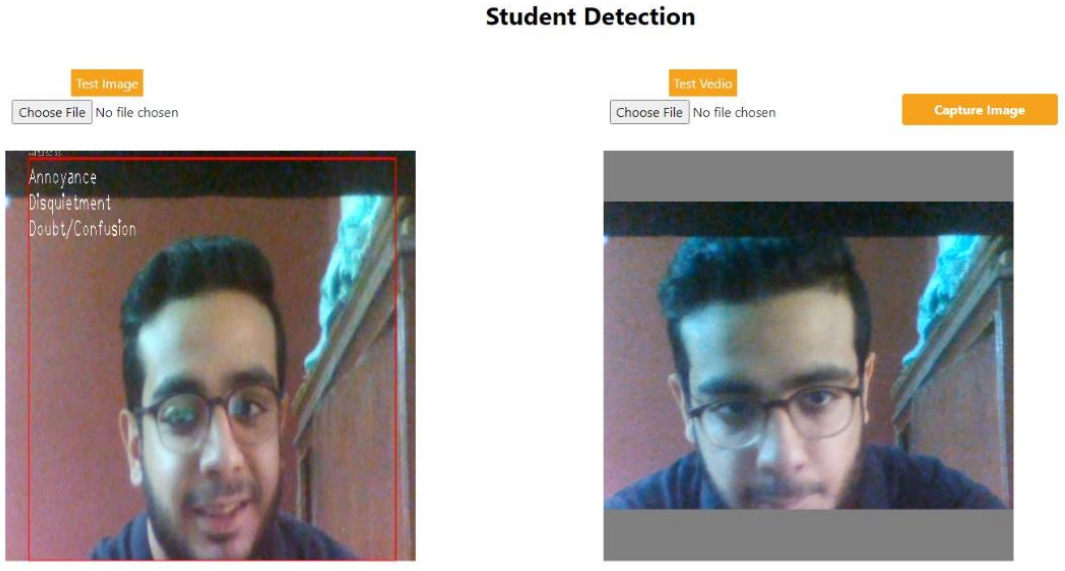


Figure 24 - User Manual Capture Image from Live Camera

2. Upload image/video button
   the user can upload image/video from its device and send it to the model to check all Persons in the image/video and get emotions for everyone.

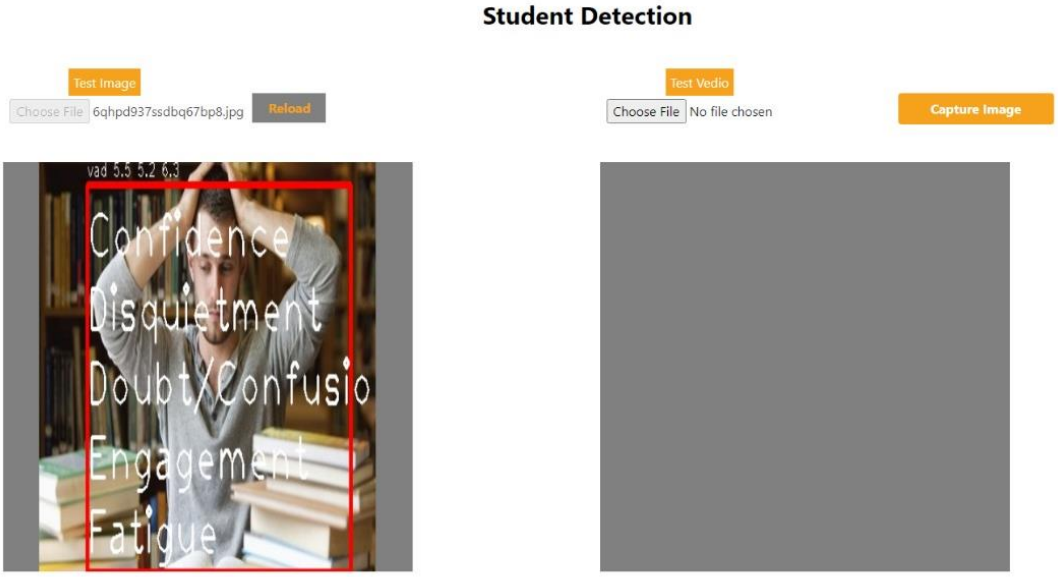


Figure 25 - User Manual Upload Image from User Device

Step (4): About US Section:

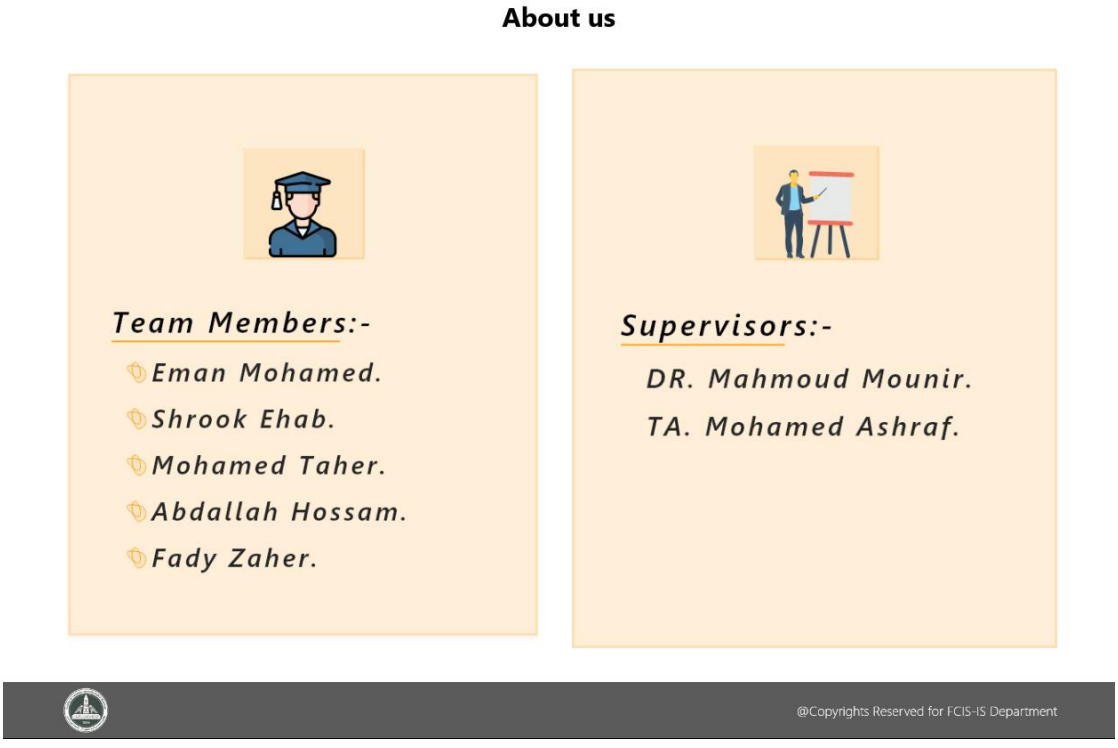


Figure 26 - User Manual About US Section

6. Results:

6.1. Project Survey:

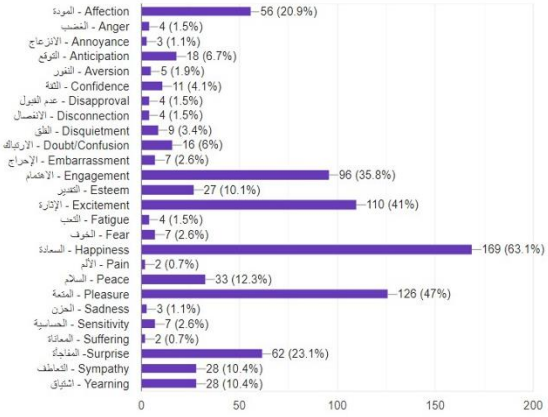


Figure 27 - Project Survey 1

In this picture, we will notice that the girl is laughing while watching the movie in the cinema

So, when we did this survey, the opinions differed regarding the description of the girl's expressions. Some of them said affection, others said interest, some said excitement, some happiness, some fun, and others many expressions as shown in (Figure 27 - Project Survey 1).

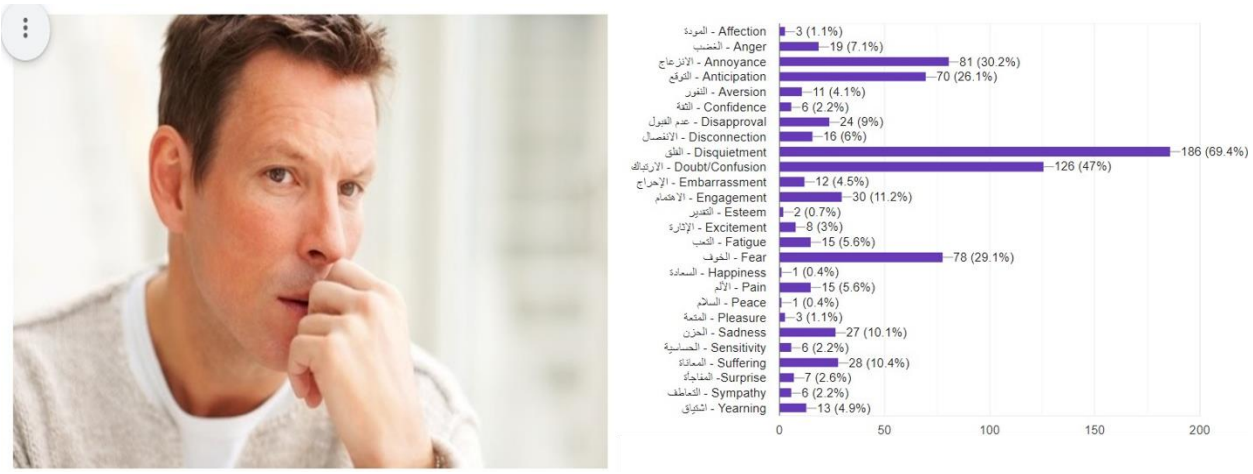


Figure 28 - Project Survey 2

In this picture, we will notice that this person is meditating and thinking about something

So, when we did this survey, opinions differed in describing the expressions of this person, and some of them said annoyance, Others said Disquietment, and some said doubt, fear, some said sadness, and many other expressions as shown in (Figure 28 - Project Survey 2).

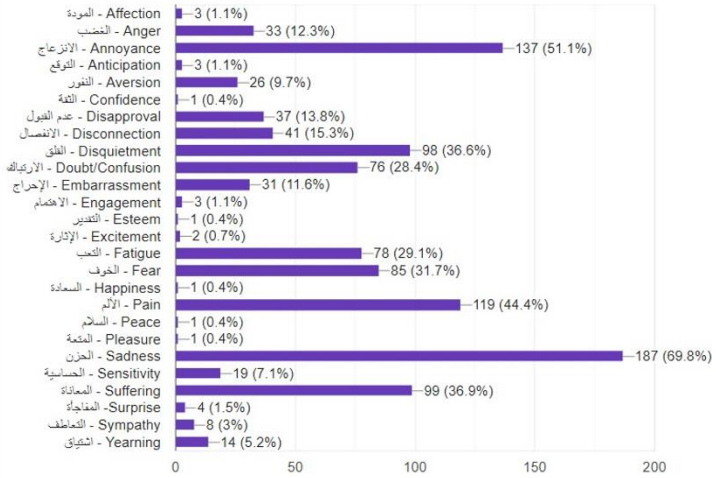


Figure 29 - Project Survey 3

In this picture, we will notice that this girl is very sad.

So, when we did the survey, opinions differed in describing the expressions of this girl. Some of them said annoyance

Others said Disquietment, and some said doubt, fear, sadness, tired, embarrassed, angry, and many other expressions, as shown in (Figure 29 - Project Survey 3).

- **The Conclusion from this survey:**

  From the opinions of people in the pictures that we presented, we noticed that people did not agree on one expression, but there were many filtered expressions

  <u>And this is what we want to prove through this survey.</u>

  The idea of our project is not easy. It does not have a single answer. Rather, there are many answers that are provided when asking about the expressions of any person. It is not easy to know the answer. It is a problem that does not have a single solution.

  It is a problem that is difficult for a person to define and there is no measure for it

  Therefore, it is difficult for the machine to determine the expressions of any person in a large proportion, and this is what we wanted to introduce to you through this survey.

6.2. Model Improvements:

After building the model using ResNet-18, we decided to increase the number of layers so instead of using 18 layers we used 50 layers deep to optimize the algorithm and improve the model's Loss Rate as shown below:

- How the model optimized when the network starts to converge?
  - When the deeper network starts to converge, a degradation problem has been exposed: with the network depth increasing, accuracy gets saturated and then degrades rapidly. Such degradation is not caused by overfitting or by adding more layers to a deep network leads to higher training error. The deterioration of training accuracy shows that not all systems are easy to optimize.

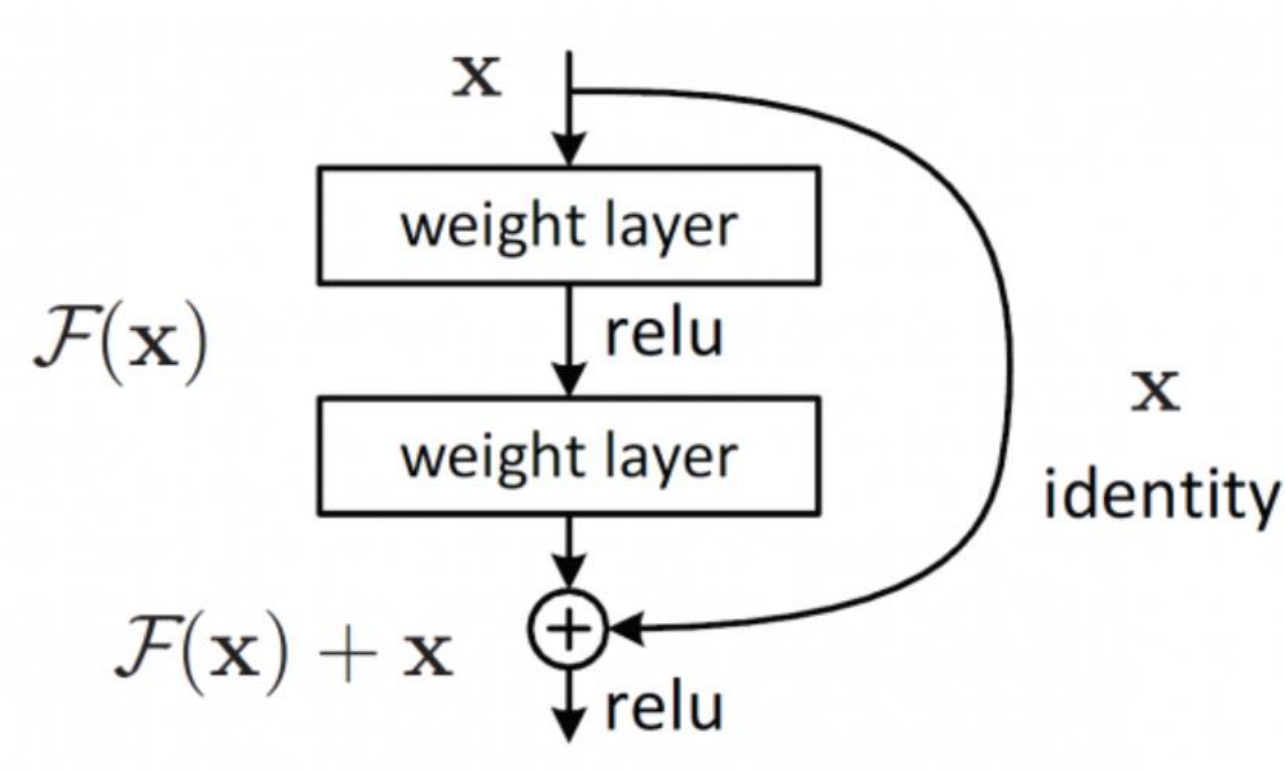


Figure 30 - ResNet Layer

  - To overcome this problem, so we use Microsoft deep residual learning framework. Instead of hoping every few stacked layers directly fit a desired

underlying mapping, as they explicitly let these layers fit a residual mapping. The formulation of F(x)+x can be realized by feedforward neural networks with shortcut connections. Shortcut connections are those skipping one or more layers shown in (Figure 30 - ResNet Layer). The shortcut connections perform identity mapping, and their outputs are added to the outputs of the stacked layers.

- By using the residual network, there are many problems which can be solved such as:
  1. ResNets are easy to optimize, but the "plain" networks (stack layers) show higher training error when the depth increases.
  2. ResNets can easily gain accuracy from greatly increased depth, producing results which are better than previous networks.
- The projection shortcut in F(x{W}+x) is used to match dimensions (done by 1×1convolutions). If the shortcuts go across feature maps of two size, it performed with a stride of 2 as shown below in (Figure 29 - Resnet 18, 50 layers).

layer name	output size	18-layer	50-layer
conv1	112×112	7×7, 64, stride 2	
conv2_x	56×56	3×3 max pool, stride 2	
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$

Figure 31 - Resnet 18, 50 layers

After building ResNet with 50 Layers we begin to compare between ResNet-18 and ResNet-50 in both (Training and Validation) phases to prove that the loss rate (error) will degree when the number of epochs is increased.

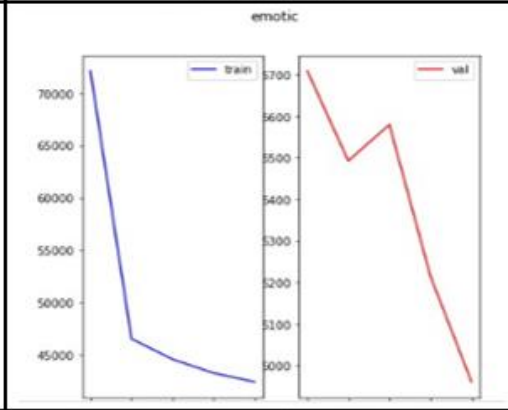
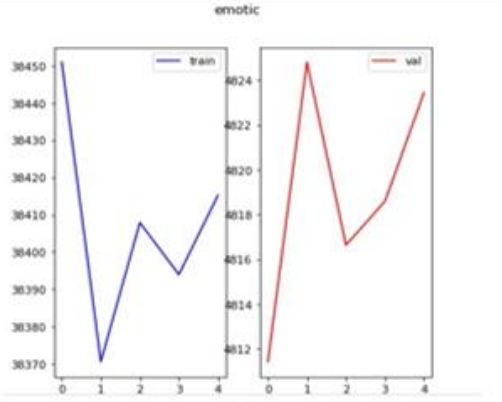
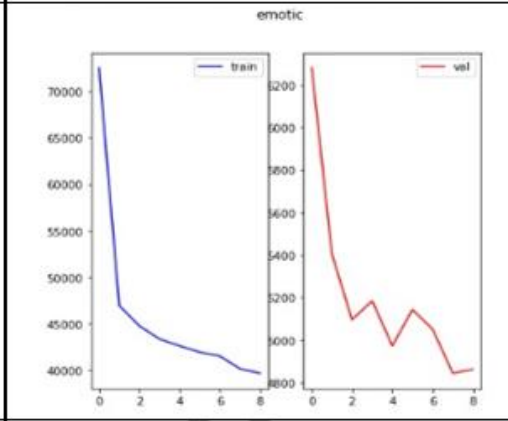
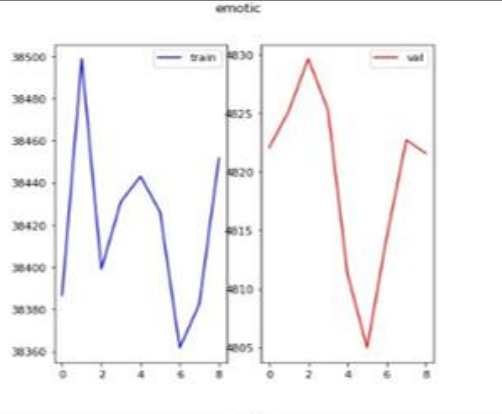
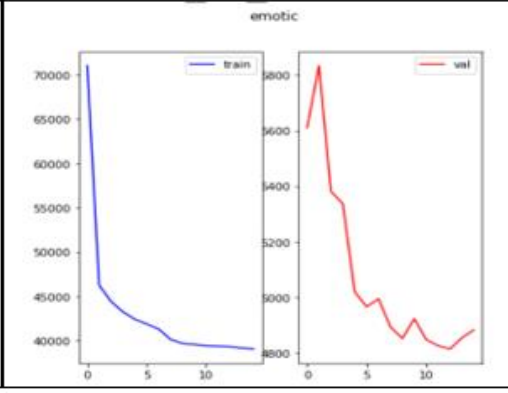
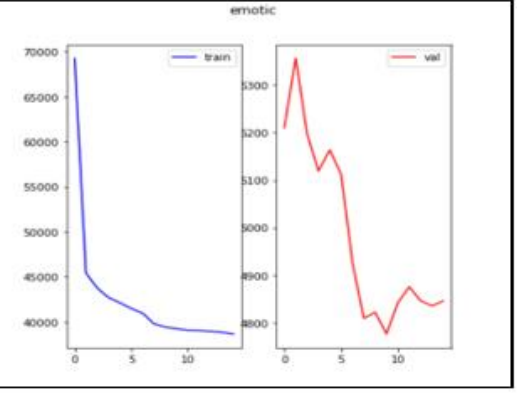
#Epochs	Resnet-18	Resnet-50
5		
9		
15		

Figure 32 - Comparison between ResNet-50 and ResNet-18 Loss

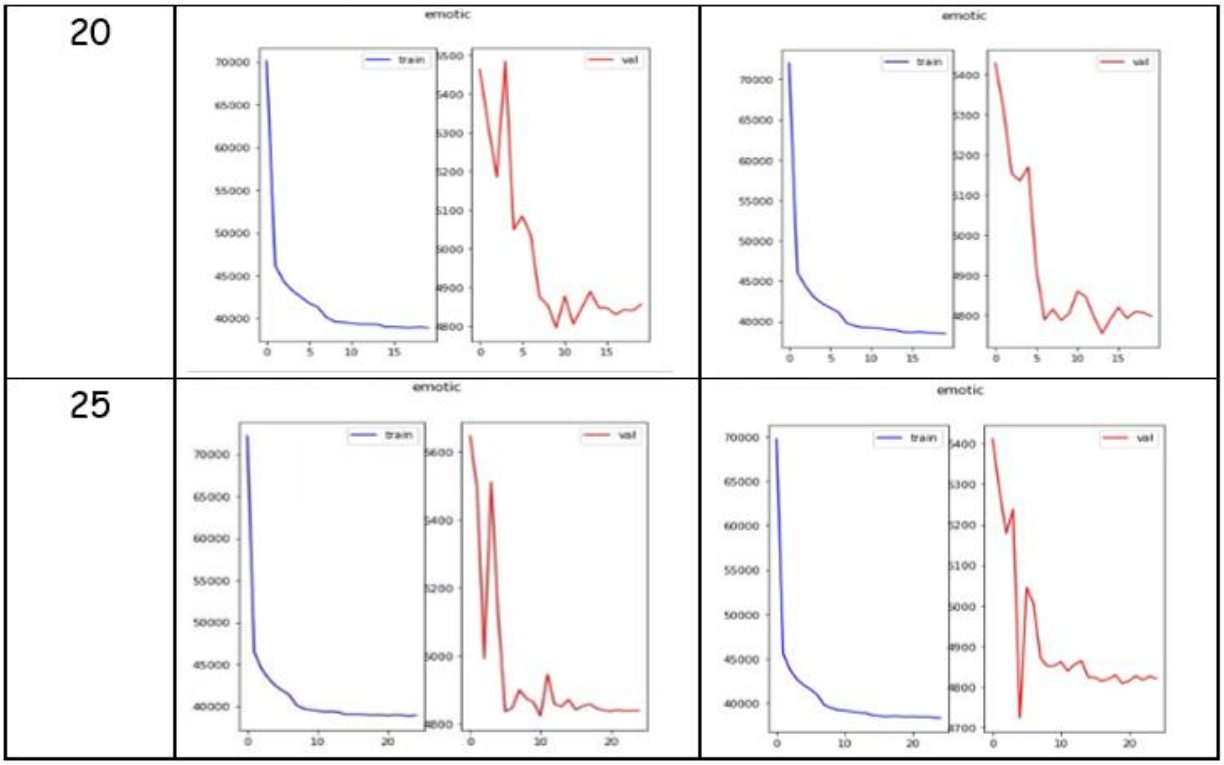


Figure 33 - Comparison between ResNet-50 and ResNet-18 Loss 2

- **The conclusion from this Comparisons:**

  From the previous 2 Figures we noticed that the loss begins to increase form **5 epochs** until it reaches **15 epochs**, after that the loss rate remains as it in both phases (Training and Validation).

  <u>And we noticed that each epoch takes about 4 minutes in (Training + Validation) phases.</u>

6.3. Model Accuracy:

When we choose to continue with ResNet-50 as it has less loss rate, we begin to compare between ResNet-18 and ResNet-50 according to specific learning rate and dropout parameters as shown below:

#Iter	Algorithm	Epochs	Training Loss	Validation Loss	Discrete Emotions (Validation)	Discrete Emotions (Test)	Accuracy (Average)
At learning rate = 0.001, Dropout = 0.5							
1.	Resnet-18	5	42407.5975	4960.6908	nan	nan	nan
2.	Resnet-18	9	39718.2243	4861.8189	nan	nan	nan
3.	Resnet-18	15	39052.7824	4883.1666	0.3412	0.2523	0.29675
4.	Resnet-18	20	38892.7058	4855.2739	0.3427	0.2530	0.29785
5.	Resnet-18	25	38920.9866	4838.7579	0.3425	0.2529	0.2977
6.	Resnet-50	5	38415.2182	4823.4588	0.3505	0.2599	0.3052
7.	Resnet-50	9	38451.7319	4821.5541	0.3505	0.2601	0.3053
8.	Resnet-50	15	38643.0324	4846.3379	0.3487	0.2586	0.30365
9.	Resnet-50	20	38487.4888	4798.1061	0.3499	0.2595	0.3047
10.	Resnet-50	25	38341.0126	4820.4284	0.3506	0.2601	0.30535
At learning rate = 0.0001, Dropout = 0.7							
11.	Resnet-50	5	50303.1504	5163.7655	0.3006	0.2279	0.26425
12.	Resnet-50	15	44528.2222	5011.1820	0.3293	0.2448	0.28705

Figure 34 - Comparison between ResNet-50 and ResNet-18 Accuracies

- Dropout: is a regularization technique for reducing overfitting in neural networks by preventing complex co-adaptations on training data. It is an efficient way of performing model averaging with neural networks. The term dilution refers to the thinning of the weights. The term dropout refers to randomly "dropping out", or omitting, units (both hidden and visible) during the training process of a network.

- learning rate (step size): is the amount that the weights are updated during training. It perhaps the most important hyperparameter, If you have time to tune only one hyperparameter. It has a small positive value, often in the range between 0.0 and 1.0 or it can be calculated by the equation shown below:

$$learning\_rate = initial\_learning\_rate * \frac{1}{1 + decay * iteration}$$

Where:

1. learning_rate is the learning rate for the current epoc.
2. initial_learning_rate is the learning rate specified as an argument to Adam Optimizer.
3. decay is the decay rate which is greater than zero.
4. iteration is the current update number.

- **The Conclusion from Using Dropout:** the accuracy of the model built using ResNet architecture increases, but not using great dropout or not using at all, as at both situations, the model, either increase the overfitting or will not learn the concepts correctly.

- **The Conclusion from Using Learning rate:** There are three learning-rate starting points to play with (i.e., 1e-1, 1e-3, and 1e-6). If the model is pre-trained, consider a low learning rate less than 1e-3 (say 1e-4). otherwise, consider a learning rate greater than or equal 1e-3. And stop learning rate when the model stops improving.

- **Finally, we choose learning rate = 1e-3 and Dropout = 0.5, which is the best accuracy we achieve using them.**

7. Conclusion and Future Work:

7.1. Conclusion:

This Website project we are using body language with facial expressions to identify the human action in natural scenes, e.g., a girl attending birthday party, a boy is talking with his friends in the park, …., etc.

But our main target is the instructors and their students, so in our project we are showing to the instructor the feeling of the student in the lecture from a **live website**, by **detecting** the facial expressions and body movements of every student whether he/she is (angry, fear, happy, …., etc.).

Also, it makes predictions not just using the body language and facial expressions, but using the context of the image which means the surrounding area that the student is exist in.

7.2. Future Work:

For now, our project has achieved a good performance compared to the previous works, but it still could be improved again to take less time in processing time.

There are features that could be added to help the users use this website easily for example:

- **Record video in live stream:**
  When the live camera is opened, the user could click to record a video not just an image.

- **Capture an image without clicking:**
  For example, each minute the system takes a snapshot and makes predictions on it then display it without the user clicks.

- **Record a video without clicking:**
  For example, each minute the system records a short video and makes predictions on it then display it without the user clicks.

- **Publish the website:**
  Publish the website on the internet so the users could use by the URL.

If these features are added, then the website will be easier to use and would be a great improvement to it.

Many users will benefit from the website after improvements due to Covid-19 disease and everyone is staying in their home and finish their works from home.

8. References:

1. Vikas Kumar, Shivansh Rao and Li Yu, (2020) “REAL TIME FACIAL EMOTION RECOGNITION WITH DEEP CONVOLUTIONALNEURAL NETWORK”, arXiv:2008.02655.
2. Panagiotis Paraskevas Filntisis, Niki Efthymiou, Gerasimos Potamianos,and Petros Maragos, (2020) “Emotion Understanding in Videos Through Body, Context, and Visual-Semantic Embedding Loss”, arXiv:2010.16396v1.
3. Trisha Mittal, Pooja Guhan ,Uttaran Bhattacharya, Rohan Chandra ,Aniket Bera ,Dinesh Manocha,(2020) “EmotiCon: Context-Aware Multimodal Emotion Recognition using Frege's Principle”, arXiv:2003.06692.
4. R. Kosti, J.M. Álvarez, A. Recasens and A. Lapedriza, (2019) "Context based emotion recognition using emotic dataset", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI).
5. ICCV 2019 , Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park , Kwanghoon Sohn, (2019) “Context-Aware Emotion Recognition Networks”, arXiv:1908.05913v1.
6. Metallinou A, Lee CC, Busso C, Carnicke S, Narayanan S (2010) The USC CreativeIT database: a multimodal database of theatrical improvisation. In: Workshop on Multimodal Corpora, LREC.
7. Hwang BW, Kim S, Lee SW (2006) A full-body gesture database for automatic gesture recognition. 7th International Conference on Automatic Face and Gesture Recognition (FGR06).
8. Gross R, Shi J (2001) The CMU motion of body (MoBo) database. Carnegie Mellon University.
9. https://en.wikipedia.org/wiki/Body_language
10. https://www.researchgate.net/publication/51780012_Emotion_Expression_in_Body_Action_and_Posture

11. https://psycnet.apa.org/record/2011-25181-001
12. https://www.redhat.com/en/topics/api/what-is-a-rest-api
13. https://en.wikipedia.org/wiki/JSON
14. https://en.wikipedia.org/wiki/JQuery
15. https://en.wikipedia.org/wiki/Bootstrap_(front-end_framework)
16. https://medium.com/beginners-guide-to-mobile-web-development/whats-new-in-css-3-dcd7fa6122e1
17. https://en.wikipedia.org/wiki/HTML5
18. https://en.wikipedia.org/wiki/PyCharm
19. https://en.wikipedia.org/wiki/Python_(programming_language)
20. https://en.wikipedia.org/wiki/Visual_Studio_Code
21. https://en.wikipedia.org/wiki/Django_(web_framework)
22. https://towardsdatascience.com/understanding-and-coding-a-resnet-in-keras-446d7ff84d33
23. https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/
24. https://www.analyticsvidhya.com/blog/2018/12/practical-guide-object-detection-yolo-framewor-python/
25. https://viso.ai/deep-learning/yolov3-overview/
26. https://medium.com/saarthi-ai/deploying-a-machine-learning-model-using-django-part-1-6c7de05c8d7
27. https://neurohive.io/en/popular-networks/resnet/
28. https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/
29. https://towardsdatascience.com/a-bunch-of-tips-and-tricks-for-training-deep-neural-networks-3ca24c31ddc8