**ASSVS: AUTONOUMOUS SURVEILLANCE SOLUTIONS FOR PUBLIC AND PRIVATE SECTORS**

# Default Research Focuses:

# Introduction

Fourth generation technological evolution is reshaping every sector of the world including industry, health, economy and societies. Emerging paradigms of research and development are transforming security and protection protocols as well across the globe. Manual security systems are being replaced by cameras and further autonomous systems. In the current era the most suitable and significant security agent are CCTV cameras which are bringing ease in society particularly in the domain of surveillance. As the human life is top priority therefore, efficient and robust surveillance systems are highly desirable at industrial, commercial and residential sites. Manual and semi-automated surveillance systems have failed to detect the calamities accurately along with timely intimation therefore, machine vision has become need of the hour to cope with security challenges and existing loopholes is such systems. Artificial intelligence is emerging as a game changer is all sectors of the world. Particularly, machine vision is the most significant leaf of AI tree, harnessing fidelity of deep learning models in order to train a machine for processing of camera streams. Combination of camera and deep learning-based video processing has originated a machine eye which is robust and accurate.

Our main target is surveillance videos generated in development countries. In contrast to highly developed countries, videos generated in Pakistan contain very dense proportion of human beings, variety of human actions and activities, amalgam of objects with uncertain properties and attributes. In addition to this, there are several image acquisitions errors that may accumulate in the captured videos such as occlusion of objects by one another, videos captured at night or under dim lights, scaling and zooming issues etc. These errors reversely affect quality of the captured videos. In this project we aim to generate smooth and human understandable description for video streams minimizing the effects of such errors. Moreover, our project is aimed at detecting malicious or criminal intent using security cameras. These deployed cameras record videos of hundreds of people daily. These videos are discarded after a short time due to lack of storage capacities. In our project, we convert these recorded videos into textual labels which are quite small in storage size so do not need to be removed. Our purpose is to use these videos to help detect and predict the possible crimes before they occur along with generating daily statistics.

All these problems can be addressed through deep learning based distributive video processing using state of the art edge computing devices. Proposed project is a site surveillance system using computer vision and machine learning techniques for automatic understanding of video streams for multiple applications. Our surveillance system will use cameras installed at required sites either remote, public and private territories including all sectors ranging from domestic to commercial and military for video stream. Machine vision based on deep learning algorithms will follow the camera stream in order to perform video processing for person detection, facial expression recognition, action recognition and anomaly detection. Deep learning algorithms will be embedded in small GPUs to make this detection and recognition process distributive rather central. Camera streams will be accessed and processed at run time using aforesaid edge computing devices transforming the existing cloud-based processing surveillance systems to on node detection and recognition systems. Our proposed distributive surveillance system will be connected with central server to intimate the central control unit about anomaly for prompt action. However, the video processing with be onsite and autonomous saving the resource sources followed by a robust and efficient surveillance protocol. Proposed system will be a cheap, compact and resource efficient approach to enhance the security of remote sites which are usually scars in resources. Canvas of proposed system will be generic covering all the possible sectors requiring security and surveillance.

**Research Questions**

1. What type of patterns and knowledge can be extracted from big datasets for Video Surveillance?
2. What type of data needs to be collected and archived for development of automatic video surveillance applications?
3. What would be the nature of developed algorithms; statistics based, infographics, intensive or textual reports related?
4. How the video stream data can be used for the tracking of persons and vehicles in public areas?
5. Hoe the video stream data can be utilized to increase the security of private places, surveillance areas and shopping malls.

# Literature Survey

The latest trend in the vision area is the integration of natural language processing. Although, the areas of computer vision like object detection, human activity recognition and human emotion detection or tending towards maturity. It is becoming popular to introduce natural language concepts into a vision system. While literature relating to object recognition, human action recognition, and emotion detection are moving towards maturity, automatic description of visual scenes is still in its infancy. Machines can detect human activity in videos to a certain level, but there are few efforts to automatically describe visual scenes. The majority of research on video retrieval has relied on keywords. Every visual sight can be divided into foreground and background, with foreground acting as the primary "region of attention" and backdrop aiding with foreground comprehension. Important visual elements like objects and their actions are frequently regarded as being in the foreground.

Since "person" is frequently the most significant and fascinating aspect of any film, a survey of approaches pertaining to human anatomy and activities is provided. Identification of humans in the stream can be aided by the existence of human faces or bodies (upper or lower, complete). Identification based on the human body or face opens up a number of intriguing uses. Gender, age, and emotional information can all be determined largely by a person's facial features. Body data is useful in identifying human behavior and how they interact with other objects. We quickly review the research on human detection using facial and physical characteristics.

## **High Level Feature detection:**

Age can be determined only from facial features. Two primary categories of features are used by one face facial age estimation algorithm: Geometrical ratios determined by the separation and size of specific facial features, as well as a calculation of the number of wrinkles found by deformable contours (snakes) in facial regions where wrinkles are typically seen. Faces were divided into categories for infants, adults, and seniors based on these attributes. Kulkarni et al [1] .propose aging pattern generator creates aging patterns for each individual in a dataset made up of face photos of each subject taken at various ages. A lower dimensional space can then be mapped onto each collection of temporal face images, which are thought of as single samples. A face that has never been seen before is inserted at various points in a pattern, and the location that minimizes the generalization error reveals the subject's age. Using manifold training, Fu and Huang portray aging trends [2]. The majority of age estimation techniques discussed in the literature make use of data from the entire face. Sue et al. instead relied on a three-level hierarchical face model [3] as the foundation for age estimate.

Only facial data can be used to determine a person's gender. Once more, processing on the face components comes first. From the frontal photographs of a face, in [4] calculated 16 geometric features, such as pupil to eye brow spacing and eye brow thickness. The gender was determined using these characteristics. Sun et al. used frontal photos to determine genetic feature subsets. Additionally suggested [5] is a multi-modal gender classification method that makes use of both voice and visuals. Using frontal faces, Mogadham and Yang studied the issue of gender identification [116]. They used many classifiers, including SVM, LDA, and ICA, to tackle this problem. For girls faces, their technique worked perfectly, but it performed less than satisfactorily for boys faces. As the male faces contain th extra features like beards and moustaches and the proposed study focused on geometric features that decreased the performance.

The most natural way that people convey their feelings is through their faces. The main challenge with categorizing facial expressions is coming up with a list of categories to work with. According to psychologists, persons have 6 core facial expressions at birth: happiness, rage, contempt, fear, amazement, and sorrow. Other expressions are picked up from their surroundings [6] Machine learning techniques or rule-based methods can both be used to identify emotions from facial expressions. Utilizing machine learning algorithms for emotion recognition. Cohen et al. [7] investigated emotion recognition from facial expression using HMM. A specific HMM that has been learned for each expression was used to simulate each facial expression. There have been six HMMs in total, each one used to depict the emotions of amazement, disgust, horror, and anger. To train a support vector machine (SVMs) predictor, example photos of faces were used, with each image labeled as belonging to a specific category of emotion [8]. The training images are displayed as points of space that have been mapped into multiple categories depending on their identities. The SVMs' trained model can then be used to foretell which class a given test input will fall within. Tong developed a hierarchical architecture built on DBNs to express probabilistic interactions between different AUs and to take into consideration the temporal changes in the development of facial actions [9].

## **Object Recognition**

The semantic divide between low-level characteristics and high-level semantic information has long been a problem for the image/video field of analytics. Concept-based retrieving, which seeks to create and make use of a collection of intermediary semantic concepts, has emerged in the last decade as a novel retrieval strategy to close this gap. In order to increase retrieval performance, Zhou et al. highlighted frequently visual content in video files [10]. The topics covered by these concepts span a wide range [11] and include those pertaining to people (face, anchor, etc.), acoustics (speech, music, significant pause), objects (image blobs, buildings, graphics), location (outdoors/indoors, cityscape, landscape, studio setting), genre (weather, financial, sports), and production (camera motion, blank frames).

Detecting semantic concepts. Not recognizing objects is not the study's main objective. Visual object recognition attempts to locate the object in the image, determine its attitude, and determine whether it is present at all. Typically, this entails creating an object representation that can simulate how an object will seem when captured under a variety of imaging circumstances, such as changing object and camera attitude, scene lighting, partial occlusion, and sometimes deformation. A strong enough depiction should be able to handle a lot of background noise.

The transformed Dirichlet method (TDP), which is created using topic models with spatial transformations and geometric requirements, allows Monte Carlo algorithms to identify items in street and office settings [12]. Unsupervised stochastic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) were used in a method to extract several object types and their positions from a batch of photographs [13]. Another hierarchical probabilistic models [12] use low level information to find and identify objects in crowded natural situations. The multiclass object classification problem was approached using a model of visual object recognition with biological influences [14]. It is also possible to recognize and locate numerous item types simultaneously using a predictive model [15].

## **Current State of the Art**

In the following table present some of the related research work which is directly related to proposed work. It is important to note that most of the previous work was very limited in scope i.e., identification of limited set of objects such as human, table, chair, road, buildings etc., human actions where background was usually static and uniform, human emotions where frontal face was quite evident or description generation using limited number of extracted HLFs.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Method*** | ***Authors*** | ***Methodology*** | ***Pros & Cons*** |
| Intelligent Video Surveillance System Architecture for Abnormal Activity Detection | M. Elarbi-Boudihir, Khalid A. Al-Shalfan  Artificial Intelligence Lab, Imam University. Riyadh. KSA. | Used SVM for modeling the human activities. | Efficient for detecting the moving objects.  Made robust calibration on data extracted from different cameras.  There is no sequential numbering of frames. |
| Human Focused Video Description | Muhammad Usman Ghani Khan, Lei Zhang, Yoshihiko Gotoh | Generated the 20 distinct annotation of hands and used them further for caption generation. | The bottom-up method was used in the study to produce a natural language definition for image sequences in a time series. High level features (HLFs) are recognized as keywords in video sequence.  Need further development for complex interactions detection. |
| Real-time Abnormal Motion Detection in Surveillance Video | Nahum Kiryati, Tammy Riklin Raviv, Yan Ivanchenko, Shay Rochel | The implementation makes advantage of the macro block motion vectors that are already produced as part of common video compression techniques. | Used the vector motion instead of pixel that reduce the data rate and allowing the real time processing of video sequence  It store the video for the later processing rather then the detection objects itself. |
| Natural Language Description of Videos | Muhammad Usman Ghani Khan | Extract the high-level features of human face for the caption generation of video frame. | Produce the description of the video and evaluate the machine generated description with the description generated by human.  Aim is to produce informative and coherent description given a video. The objective of this evaluation is to measure the information content of machine generated descriptions using human authored descriptions as a reference.  The fact that unit-based phrases might not be adequately structured by merely grouping them together presents a possible issue. These phrases also have other problems, such as redundancy and verbosity. |
| Human Activity Detection and Recognition for Video Surveillance | Wei Niu, Jiao Long, Dan Han, and Yuan-Fang Wang | HMM was used for the activity recognition by utilizing the simple statistical features of trajectories | Frame process of separating and feature correlation are examples of low-level motion detection algorithms that are the foundation for intelligent control and fail-over mechanisms for activity tracking and detection.  Problems relate to 2) camera calibration 3) also ignore the processing of sub pixels to gain high accuracy. |

# Methodology

The study presented in this research work is concerned with classification, detection and recognition of objects and activates for visual images in a time series using a bottom-up method. Primarily high-level features (HLFs) are recognized from individual video frames. They may be ‘keywords’, such as a particular object, its position/moves and its interaction with other objects in the video. Spatial relations between HLFs are calculated to improve the explanation of the semantics of visual scene. Extracted HLFs and spatial relations between them are then utilized to provide solutions to multiple applications.

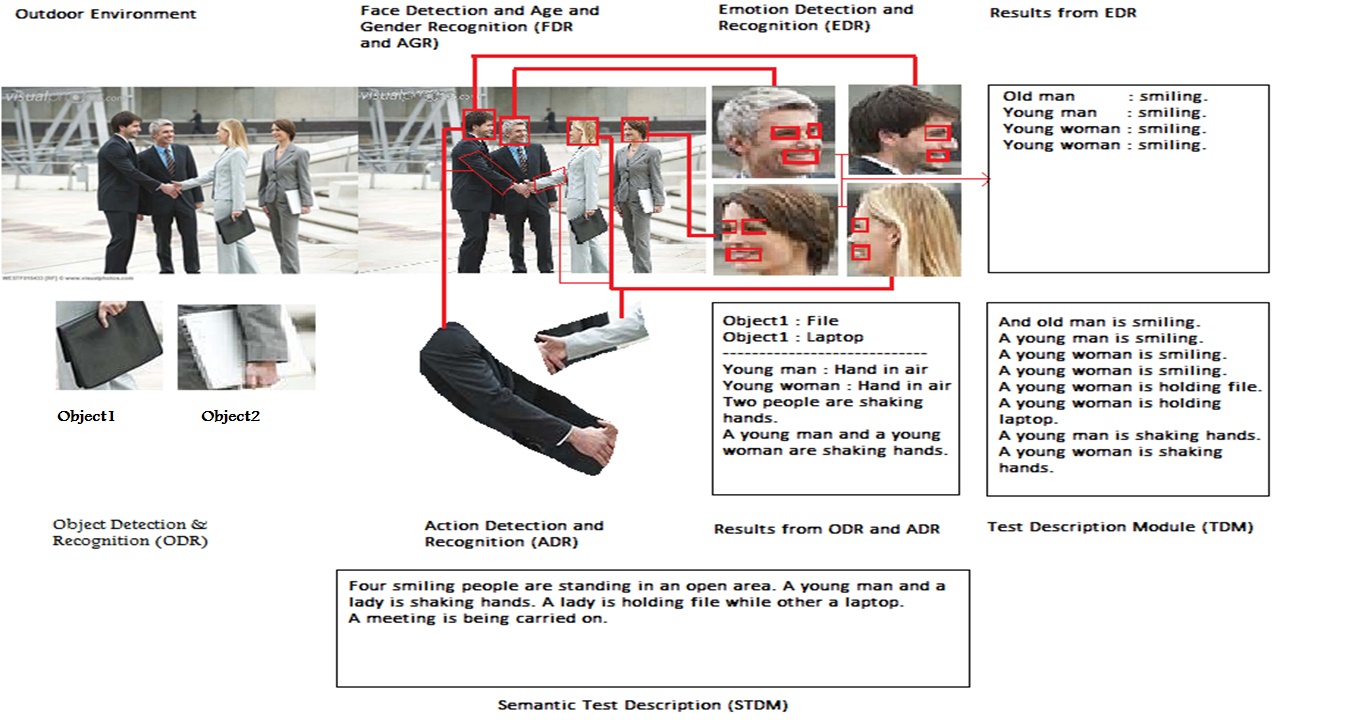
In the following, we describe details of each of the above-mentioned modules.

## **Processing High Level Features:**

The most important aspect of a video stream is the presence of humans in it and his interaction with surrounding objects. Presence of human and his gender, emotions, expressions, postures and interactive objectives bring a lot of information about the scene statistics. Following list summarizes the features of this part of the project:

1. Age/Gender Recognition
2. Emotion Recognition
3. Activity/Action Recognition
4. Face Recognition
5. Human Object Interaction

Details of each of these tasks are described below figure.



## **Human and Object Identification:**

Identification of human face or body can prove the presence of human in a video. A faster region based convolutional neural network (FRCNN) [8] can be employed to localize human and other objects of interest in an image. The FRCNN consists of multiple networks, backbone network and the regional proposal network and classification and box regressor network. The test image is scanned and proposed foreground regions are passed to the classifier that outputs the class probability and boundary box regressor outputs the bounding boxes of objects. This way multiple objects along with human can be identified from one image.

## **Human Age/ Gender Detection:**

Human face is of fundamental importance for guessing age of a person. Facial features play an important role in identifying age, gender and emotion information. Deep learning-based classifiers have this tendency to extract edge, color and geometric features automatically. We aim to develop a robust classification network that can identify human age and gender from facial image. The classification network will have two output layers one for age class probabilities and one for gender probabilities. For this work, we limit our scope to classify humans in three age groups; baby, young and old aged.

## **Emotion Recognition:**

Emotion recognition comprises of three steps; face detection, facial features localization, and emotion classification. The features that are commonly used to characterize and define a human face are the eyes and the mouth. Convolutional neural network is trained for seven emotion classes neutral, smile, sad, anger, surprise, fear and cry [10]. Confusion matrix can provide a best mechanism for evaluation results of this part.

## **Activity Recognition:**

The task of activity recognition is categorized into three levels for representation as described in figure 20. The first level is the basic or core level to extract the actions or activity from the video sequence. It involved the segmentation of human object from background, then, feature extraction of human silhouette. Subsequently, activity detection is applied to recognize the action or activity on the basis of these features. Mid-level representation is about to justify that a single or multiple persons are involved in activity. Then on the basis of recognized actions and scene understanding, it is identified that whether the activity is normal or not. Last, the third level represents pragmatic application for this activity recognition system.

State of the art 3D- convolutional neural networks tackles the patterns from spatial and temporal data. Such set of convolutional layers can be employed to extract robust features from videos and finally a set of layers to classify the activities into one of the target classes employed.

## **Face Recognition:**

We propose the methodology based on the Faster RCNN, which is the latest deep learning technique. This makes our proposed algorithm efficient in terms of speed and accuracy. In Faster RCNN, we have two different networks. The first one is the region proposal network and the second one is the backbone architecture. In the face detection and recognition system, we have to store the convolution features of the specific ROI, in which the face has been detected. So, in case of face recognition, we extract the 128 features from the last fully connected layers of our proposed architecture. So, whenever the query face arrives, the algorithm after face detection matches the 128 convolution features with the features which have already been stored in the database.

## **Face mask detection:**

Face mask detection helps in monitoring the observation of SOPs imposed by the government to prevent the spread of COVID-19. Deep learning-based techniques will be employed to detect the person with and without mask. Region based convolutional neural network is developed to identify the persons wearing masks from real-time camera streams.

The aforementioned modules are assembled with web-services and on-node embedded devices to provide solutions for following sectors

* Industrial sectors
* Offices
* Educational institutes
* Remote Areas

The products that are developed for these sectors are described below

1. Smart Pass
2. Heavy mechanical industry surveillance
3. Remote areas surveillance
4. Footfall Monitoring

Following sections give the brief description of these products

## **Smart Pass**

Smart Pass enables an organization to monitor and to track the health status of your employee. Following are the features of Smart Pass.

* Face recognition
* Mask detection
* Temperature Monitoring
* Social Distancing
* Area Occupancy
* COVID-19 symptoms detection

The above-mentioned features can be provided using on-node embedded device such as Jetson Nano as well as can be provided as backend services.

## **Heavy mechanical industry surveillance**

The product aims to detect the alarming situations that often occur in heavy mechanical industry such as fire, smoke and movement of employees without safety protocols. Safety protocols include wearing helmets, safety shoes/boots and uniform. Following are the features of the heavy mechanical surveillance product.

* Fire and Smoke Detection
* Safety Helmet Detection
* Uniform Detection
* Safety Boot Detection

Reports/Visibility:

* Live Feed
* Real-time reports
* Dashboard: (Snapshot appended)
* Alarms for deviations/incidents
* Single click on alarms to move to live feeds of deviations
* Responding/Not Responding status of the camera

The product is deployed on CCTV camera streams and computations can be performed both on Jetson Nano or as web services on the central processing server.

## **Remote areas surveillance**

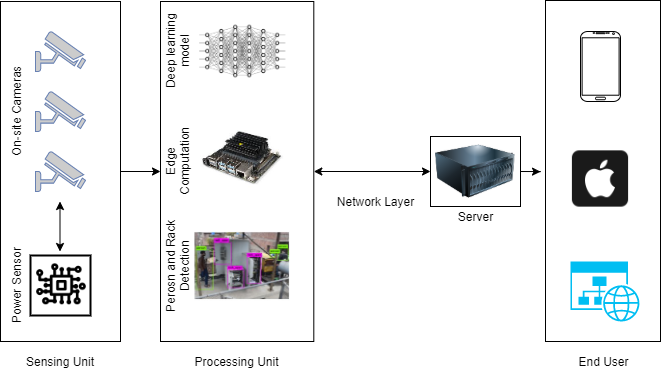
This project deploys a Deep learning-based tower surveillance system in a very compact shape. Overall system consists of an IP camera for real-time streaming, Nvidia Jetson Nano module for inference, indicators, power fail intimation circuit and a server connected with Android/IOS/Web based application for site intimation along with its video stream. Our system will provide a fully-automated security and monitoring system for remote sites without the need of any kind of manpower or guards etc. for continuous monitoring of camera feeds. Upon any kind of unusual situation, the concerned person will be notified at real-time over the internet. On node deep learning research paradigm in not very much explored right now. Site implementation of this research direction is not much developed and out proposed system is state of the art in this aspect.

Figure 1 Tower Surveillance System Diagram

Our proposed system has five modules including sensing unit, processing unit, network layer, server and end user. The system starts from the sensing point to end user in order to ensure reporting and detection based on the feeds of one or more sensing cameras as presented in figure 1.

## **Footfall Monitoring**

Footfall monitoring follows the same idea of object detection in order to detect person. In addition to person detection, person tracking algorithm is also employed in order to avoid recursive statistics. Person tracking is performed by feature mapping as well as tracking the movement of person ROI. Footfall monitoring is combined with other classification networks to identify gender, presence of mask and age of the person. The person record is stored in SQL databases and is displayed in the form of statistical reports and charts to the concerned authorities.

Monitor:

* The number of people coming in and out of a room, hall or a corridor.
* The number of occupants inside a room, hall amongst the occupants.

Reports/Visibility:

* Live Feed
* Real-time reports
* Dashboard: (Snapshot appended)
* Alarms for deviations/incidents
* Single click on alarms to move to live feeds of deviations
* Current footfall inside the premises
* Responding/Not Responding status of the camera

# Outcome and Challenges

Developing a Video understanding system for a developing country like Pakistan can be very challenging. Many things need to be taken into account both during the development of such a system and then when it is actually deployed.

The demand for intelligent automatic annotation solutions for searching, archiving, and retrieval was driven by the quickly expanding number of video collections, particularly online. All different kinds of manually labeled videos can be found in these video collections. Use some keywords to search; only a subset of movies that truly contain the intended meaning will be returned, as this annotation is typically vague, partial, and contains misspelled terms. For better indexing and retrieval, this annotation therefore has to be filtered, expanded, and validated. Moreover, consider the scenario in which a person is given a lengthy movie with a duration of hundreds of minutes and asked to examine it for the detection of suspicious activities. Even the most active people, according to a survey, struggle to effectively assess long videos. Development of such system include following challenges.

|  |  |  |  |
| --- | --- | --- | --- |
| Modality Type | Challenges | Description | Pictorial Interpretation |
| **Visual** | *Viewpoint variation* | Viewpoint variation can cause increase in the error. To ensure the robustness, model should be trained on various viewpoints. |  |
| *Illumination* | Light intensity in the images also influences the model performance |  |
| *Occlusion* | Object under consideration is being blocked by another object |  |
| *Scale Variation* | Scale variations of the objects used for processing also has an effect on model performance |  |
| *Processing Speed* | Deep learning models need high processing power; thus, speed will be low |  |
|  | *Deformation and Background Clutter* | Deformation of the object and similar background also effects performance |  |
|  | *Variance in Emotions presentation* | Some emotions are very obvious but their presentations vary from person to person. For example, a person in shock might open his mouth in curve for expression while some might not do so. |  |
|  | Clothes Impact in Detection and Recognition | But if face is covered by some clothe then its recognition is a challenge. Any part of body covered in clothes will be different to discover. | Clothing Detection for Fashion Recommendation |
| **Others** | Real Time Data Processing | Real time data processing is very difficult task and needs very challenging parameters setting. Most of the current image processing methods work offline, i.e. recorded videos are processed to find important contents. |  |
| *Security* | Since the proposed system is to be installed in sensitive areas, security can be an issue. |  |

# Research Significance

Crime identification and its prevention are major tasks in surveillacne systems that law enforcement departments have to perform on daily basis, which can be used to analyze information patterns and their relations. The manual information gathering and its mapping to the crime scene can be a clumsy and time-consuming process. Current practices are limited in terms of information gathering and in addition, may incur biases in the analysis. Moroever, the proposed solutions not only restricted to law enforcemnt agencies. One can use it for their personal business surveillance or some site surveillance etc. These solutions also includes contact less person behaviour monitoring system along with the face based attendance system. So, these solutions are also helpful to tackle the covid19 problem with the artificial intelligence.

As mentoned abobe the proposed surveillance solution is suitbael choice for every domain therefore, it is equally significant for health sector. Current ongoing pendamic situation require more caution on rushy places where implementation of Standard Operating Procedures is the top priority. The most innovative approach to tackle the covid19 or any other disease which is contegios is camera based surveaillance and intimation. Apart from health road side serveillance and smart cities are also the potential consumers of our porposed system. Potencial customer of this project will be the safe city authorities and law enforcement agencies in Pakistan. Punjab Safe City Autority Pakistan have placed cameras in several large cities of the Punjab Province Pakistan. They hired huge human force to monitor the stream of these cameras which is not still accurate enough to deal with seciruty threats effeciently. The proposed work provides them efficient and fully automatic system capable of dealing with every possible security threat without any kidn of intrupption. It will help them to reduce and eliminate the human dependency. The system automatically detects the any un-usual happening along with suspicious activities and then respond to the concerned authority for decision making. Furthermore, there are following areas where we apply our proposed solutions

1. Law enforcement agencies
2. Safe City Authorities
3. Shopping Mall Surveillance
4. Different institutional organizations
5. High Security Zones
6. Hospitals/ Bus and Train Stations
7. Airports
8. Offices

Warehouse protection is another promising sector that will require our proposed surveillance system. Here our proposed system will prevent the warehouse thefts and enhance the safety of the staff managing such mammoth places. Proposed device will authorize the entering personals followed by anomaly detection in case of authorization persons. Abnormal activities will be reported to the warehouse managing authorities for reaction against the detection action. Apart from it the device will also report hazardous event in the warehouse to keep the safety standards of workers intact. There are lot of sectors facing issues of security which can addressed using our proposed solutions and products. This includes the automation of the already existing manual attendance system, video surveillance system along with the personal behaviour monitoring system. This has also direct impact on the econmy of the country because our automatic system is cheaper and accurate to exisiting surverillance systems.

# References

[1] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, “Composing Simple Image Descriptions using Web-scale N-grams.” Association for Computational Linguistics, pp. 220–228, 2011. Accessed: Oct. 14, 2022. [Online]. Available: https://aclanthology.org/W11-0326

[2] Y. Fu and T. S. Huang, “Human age estimation with regression on discriminative aging manifold,” *IEEE Trans Multimedia*, vol. 10, no. 4, pp. 578–584, Jun. 2008, doi: 10.1109/TMM.2008.921847.

[3] J. Suo, T. Wu, S. Zhu, S. Shan, X. Chen, and W. Gao, “Design sparse features for age estimation using hierarchical face model,” *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008*, 2008, doi: 10.1109/AFGR.2008.4813314.

[4] Z. Sun, G. Bebis, X. Yuan, and S. J. Louis, “Genetic feature subset selection for gender classification: A comparison study,” *Proceedings of IEEE Workshop on Applications of Computer Vision*, vol. 2002-January, pp. 165–170, 2002, doi: 10.1109/ACV.2002.1182176.

[5] L. Walavalkar, M. Yeasin, A. Narasimhamurthy, and R. Sharma, “SUPPORT VECTOR LEARNING FOR GENDER CLASSIFICATION USING AUDIO AND VISUAL CUES,” *http://dx.doi.org/10.1142/S0218001403002447*, vol. 17, no. 3, pp. 417–439, Nov. 2011, doi: 10.1142/S0218001403002447.

[6] IEEE Communications Society., Lasers and Electro-optics Society (Institute of Electrical and Electronics Engineers), and Optical Society of America., “A New Optimization Algorithm on Pulse-width Compression Using Polarization Controls,” *Optical Fiber Communication Conference (2003), paper MF15*, p. MF15, Mar. 2003, Accessed: Oct. 14, 2022. [Online]. Available: https://opg.optica.org/abstract.cfm?uri=OFC-2003-MF15

[7] I. Cohen, A. Garg, and T. S. Huang, “Emotion Recognition from Facial Expressions using Multilevel HMM”.

[8] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan, “Machine learning methods for fully automatic recognition of facial expressions and facial actions,” *Conf Proc IEEE Int Conf Syst Man Cybern*, vol. 1, pp. 592–597, 2004, doi: 10.1109/ICSMC.2004.1398364.

[9] Y. Tong, S. Member, W. Liao, Q. Ji, and S. Member, “Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships”, doi: 10.1109/TPAMI.2007.1094.

[10] X. Zhou, X. Huang, B. Xu, and Y. Wang, “Real-time facial expression recognition based on boosted embedded hidden Markov model,” *Proceedings - Third International Conference on Image and Graphics*, pp. 290–293, 2004, doi: 10.1109/ICIG.2004.118.

[11] S.-F. CHANG *et al.*, “Columbia University/VIREO-cityU/IRIT TRECVID2008 high-level feature extraction and interactive video search,” *TREC Video Retrieval Evaluation, TRECVID 2008, Columbia University, October 26*, Jan. 2008, Accessed: Oct. 14, 2022. [Online]. Available: https://ink.library.smu.edu.sg/sis\_research/6523

[12] E. B. Sudderth *et al.*, “Describing Visual Scenes Using Transformed Objects and Parts”, doi: 10.1007/s11263-007-0069-5.

[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, “Discovering objects and their location in images”.

[14] J. Mutch and D. G. Lowe, “Multiclass Object Recognition with Sparse, Localized Features”.

[15] K. Mikolajczyk, B. Leibe, and B. Schiele, “Multiple object class detection with a generative model,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 26–33, 2006, doi: 10.1109/CVPR.2006.202.

# Project Time Table:

|  |  |  |
| --- | --- | --- |
| Task Name | Duration | Outcome |
| **Smart City and Policy Governance based on Multi-Sensor Data Analysis.** | **2 years** |  |
| **Literature Review** | **100 days** |  |
| Smart City and Policy Governance based on Multi-Sensor Data Analysis. | 40 days | Literature Survey |
| Study and research of detection ideas and methodology | 30 days |
| Smart devices and sensors methodology designing | 30 days |
| Research Paper 1 |  | Paper 1 |
| **Data Acquisition** | **115 days** |  |
| Data Collection for questionnaires designing | 25 days | Data Collection |
| Data collection from sensor data | 30 days |
| Data collection from video data | 30 days |
| Data collection from previous historical data | 30 days |
| Research Paper 2 |  | Paper 2 |
| **Dataset Annotation** | **57 days** |  |
| Images selection | 17 days | Data Labeling |
| Annotating sensor images | 20 days |
| Annotating video data | 20 days |
| Research Paper 3 |  | Paper 3 |
| **Methodology** | **130 days** |  |
| Algorithm development for extracting incongruity features | 50 days | Project Development  Paper 4 |
| Algorithm development for extracting figurative features | 40 days |
| Research Paper 4 |  |
| Development on data augmentation | 40 days |
| Research Paper 5 |  | Paper 5 |
| **Integration** | **90 days** |  |
| Integration of CNN+LSTM, Mask-R-CNN, EfficentNet and BERT | 50 days | Deployment  Paper 6 |
| Research Paper 6 |  |
| Testing and optimization | 30 days |
| System Deployment | 10 days |
| **Result Compilation** | **36 days** |  |
| Report compilation and technical writing | 36 days | Results. |
| Research Paper 7 |  | Paper |

**Funds needed**

|  |  |  |
| --- | --- | --- |
| Staff | Per Months’ Salary (Saudi Riyal) | 24 Months’ Salary (Riyal) |
| ??? (PI) |  |  |
| Prof Dr M. Usman Ghani (Co-PI- UET Lahore Pakistan) | 6000 | 144000 |
| Senior ML Researcher | 3300 | 79200 |
| Junior ML Researcher | 1650 | 39600 |
| Data Annotators | 1000 | 24000 |
| Web Developer | 1000 | 24000 |
| GUI Designer | 1000 | 24000 |
| Interns | 1000 | 24000 |
| Software Engineer- Server implementation | 1650 | 24000 |
| Data Scientist | 2000 | 24000 |
| Content Writer | 1000 | 24000 |
|  |  |  |
| Storage Space on Cloud |  | 5000 |
| Hardware model design |  | 7000 |
| Sensor’s integration |  | 8000 |
| Google Collab professional Account |  | 5000 |
| Total Project Cost (HR + Hardware) |  | 455800 |

**GANTT CHART**

