

Video Analytics Framework for Automated Parking

M. Rizwan¹, H. A. Habib²

^{1,2}Department of Computer Engineering, University of Engineering and Technology, Taxila-Pakistan

²adnan.habib@uettaxila.edu.pk

Abstract-In this research work we proposed secure parking framework based on video analytics for human activity recognition and user interaction through smartphone application. We used machine learning algorithms for human activity recognition using smart camera. Algorithms were used to process the parking lot video to extract meanings form human activities while mobile phone application was used to communicate with the user. Image registration algorithm marked incoming and leaving vehicles to keep track of available parking bins. Activities happening in the parking area e.g. human-vehicle interaction, human-human interaction, parking, entering and exiting the parking lot were recognized using support vector machine classification on space time features representation of the scene. In order to train and test the activity identification algorithm we used real world video dataset VIRAT (1.0). Video analytics algorithms were embedded into smart camera hardware. Video backup was maintained at backend surveillance server.

Keywords-Smart Camera, Video Analytics, Vehicle Surveillance, Human Activity Analysis

I. INTRODUCTION

Security and surveillance of parking lot is common requirement in urbanized areas. Various automated solutions have been proposed and installed in parking lots. These parking lot management solutions are based on motion sensors, acoustic sensors, passive infrared

sensors and radar sensors technology [i]. Video surveillance cameras are installed to monitor these parking lots. Human operators keep watch on parking lot through video surveillance console and look for activities happening in the parking lot. This paper presents framework based on video analytics in smart camera and mobile phone application for security and surveillance of vehicles in parking lots as shown in Fig. 1.

The video processing system proposed in this research consist of smart cameras, mobile phone applications and videos surveillance servers. A smart camera has its own processing unit built into the camera module to perform video analysis. Furthermore these cameras have communication capabilities to interact with video server as well. The processor in smart camera can be programmed on different algorithms for analysis of visual data. A video server stores the video archive and the human activity models as back end repository. Smart cameras are major advancement in camera technology. Vision based parking management is most convenient choice for parking management as we usually have video surveillance infrastructure in almost every urbanized parking facility. In contrast with mounted sensor array based systems video based system is more flexible and easy to scale. It can be modified to fit in various geometrical configurations and environmental conditions [ii]. Although, advantages of using video sensing are evident there are some inherent challenges attached to use of cameras. Changing lighting conditions, inclement weather, fog, shadows, vehicle projection and contrast are a few of

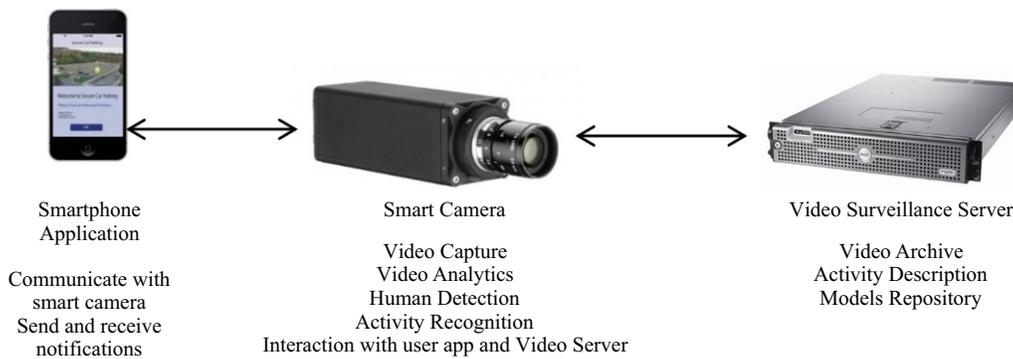


Fig. 1. System diagram of the proposed framework with details of modules

many challenges that can affect the performance of detection process. Camera motions under strong winds can also affect the performance. The geometry of the parking area can also pose a challenging effect (Fig. 2) and cause occlusion leading to misclassification of events.



Fig. 2. Examples realistic parking scenes from VIRAT dataset reference [xx] depicting light variations, shadow and projection variations.

Upon arrival at parking lot the incoming vehicle is monitored and registered to the vehicle database. The smart camera sends notification to the user smart phone app indicating the nearest available parking bin (Fig. 3).



Fig. 3. Smartphone app indicating first available parking bin for driver of the vehicle.

This research work presents a novel approach for parking lot management using video analytics for human action and interaction. This work applies region localization feature extraction mechanism in order to simplify the future processing. For activity detection, BoW representation of STIP descriptors has been used in SVM based multi classifier. This mechanism is less complex and requires minimal classifier training. This

method works directly on presented features and generates activity classes. For evaluation of proposed techniques we used VIART large scale activity dataset. VIRAT data set contains videos depicting real world environmental constraints the nature of activities in these videos makes this data set extremely challenging. For validation of obtained results “leave one out cross validation” scheme was used and classification accuracy up to 91% was achieved. In next section we have mentioned some work related to incorporation of video analytics for parking lot management.

II. LITERATURE REVIEW

Video analytics for activity recognition is the process of automatically identifying the activities of interest in a video segment. By definition high-level or complex events are long-term object interactions that happen under certain spatially and temporally dynamic settings in a scene. Usually complex activities are categorized into two classes instructional and social activities [iii]. The former includes procedural videos (e.g., “parking a vehicle”, “changing a vehicle tire”, “opening a vehicle trunk”), while the later includes social activities (e.g., “birthday party”, “conversation”).

One class of methods makes use of motion trajectories to represent actions based of target tracking as in [iv] and [v]. Another type of approaches uses background subtraction to obtain a sequence body contours to model actions [vi]. In more recent times action recognition is carried out using local spatio-temporal features which are computed over the detected spatio-temporal interest points (STIPs). These features were used to characterize the video sequence and the classification was carried out using a bag-of-word (BoW) approach [vii]. Methods for motion segmentation were also used before local feature based methods [iii].

A. Visual Feature Representation

Features are the backbone of visual content analysis. A good feature presentation is supposed to be robust against light and scale variations. This makes possible the recognition of same class of video possible under different conditions. We have two vital sources of information that can be utilized in this process. First one is visual appearance information; this is about the objects present in the scene and scene settings. The second one is the motion information pertaining to the mobility of the objects and camera.

Feature representations like scale invariant features transform (SIFT) and histogram of oriented gradient (HOG) are known as 2D feature representation schemes. They are easy to compute and have low computational complexity [viii]. These features have proven remarkably successful in presenting distinguishable visual discriminations in videos.

Their accuracy further improves in the videos in which we do not have rapid inter frame transitions [iii]. Parking lot videos are perfectly suitable to be categorized as this type of videos thus the HOG feature extractors are excellent choice as feature descriptors. For videos with rapid inter frame transition it is required to pick selective frames out of overall video sequence if not all frames are to be used. An optimal key frame selection criterion is yet to be discovered. It is usually the practice in research to pick key frames uniformly [iv]. Since these features include no

temporal information they are unable to provide any motion information which was a very important requirement in video analysis. These lead scientists to device a spatio-temporal feature representation like [ix]. The work of [v] presented the use of local and global reference points to model the motion of dense trajectories, this lead to a comprehensive representation of location and motion. This representation proved to be a more robust, and also was able to infer the relationships among moving objects

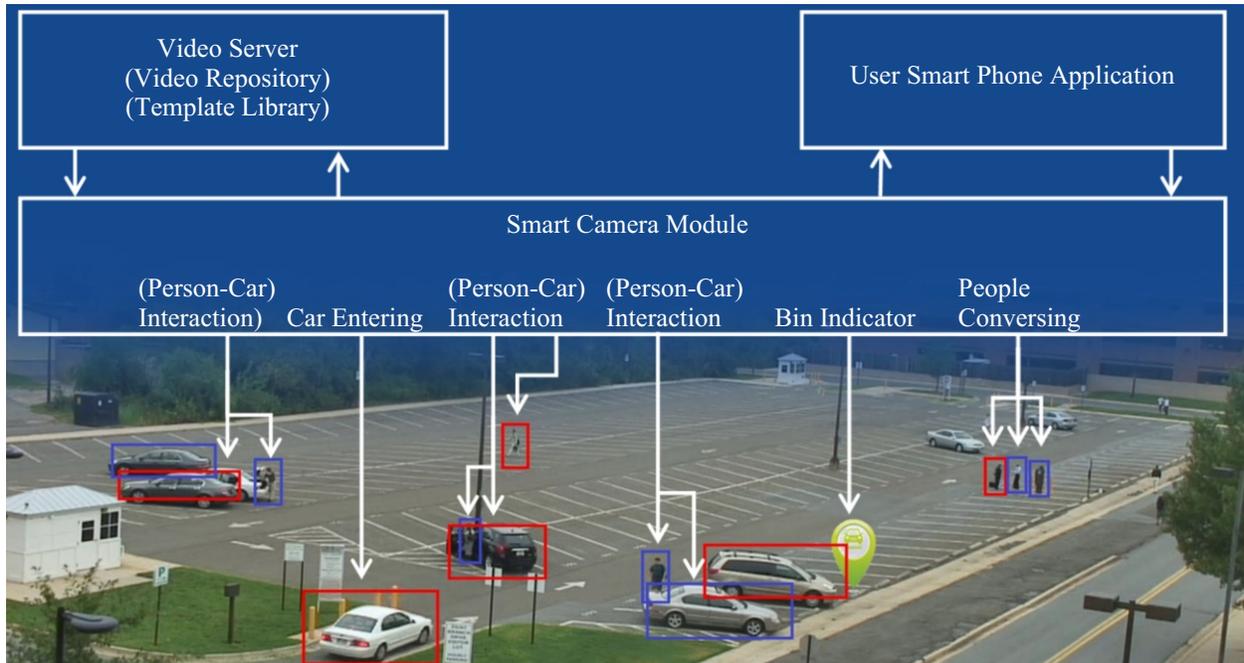


Fig. 4. The proposed activity recognition framework in process. Smart camera registers parking lot and sends captured videos to surveillance server. Smart phone app receives notification from smart camera. Human activities are analyzed in smart camera processing unit.

B. Activity Recognition methods

Once the scene is represented using features activity recognition process can start. There are various classifiers people have used in order to identify activities in videos. Activity classification is a typical machine learning process. A collection of training videos were used for model training and test videos for model verification. There were direct classification methods as well but their ability to classify highly structured activities was limited. Process of understanding deep semantic structure present in complex activities requires more structured approach [x]. For example the event “changing a vehicle tire”, consists of semantically base level categories like as “holding a tire”, “person using wrench” and “person jacking car”. A bag of words representation breakdowns information into a feature vector direct classification carried out over this does not explain this semantic structure. This lead the researchers to the discovery of more efficient semantic analyzer for

complex activity classifier. These models were successfully applied for the recognition of complex activity recognition. Reference [xi] proposed the usage of syntactic context free grammar SCFG based scene representation approach for complex event recognition. This approach was used in integration with a real time activity monitoring system to demonstrate its usefulness. In a similar manner, [xii] came up with the use of CFG, to represent an event as time based processes consisting of poses, gestures, and sub-events. A slight modification of this idea was used by [xiii]. They attempted to detect events happening in parking lot using attribute grammar posing additional conditions upon existing production rules

III. MATERIALS AND METHODS

As depicted in Fig. 4 the proposed video analytics infrastructure used smart camera for video processing. The input video frames were processed every time

there was a change in visual information in the scene. The object detection algorithms were used to detect the objects of interest and marked the areas in video for further processing. The activity recognition algorithms were used to process the given regions to identify types of activities found (Fig. 5). Following is the detailed discussion on algorithmic approach used in this paper.

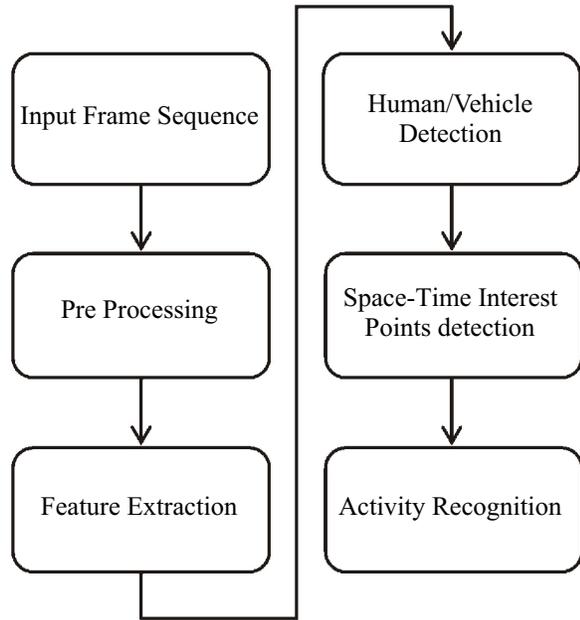


Fig. 5. Process diagram of the activity recognition method

A. Foreground Segmentation

In order to acquire foreground presentation of the object background subtraction approach was used. As the system used current frame I and subtract background frame B from it then pixel by pixel threshold is used to determine a pixel's being background or foreground as in Eq.(1).

$$|I_i(x,y) - B_i(x,y)| > T \tag{1}$$

The threshold T is predefined threshold value. The result of background subtraction is a binary map that shows presence of foreground objects in terms of region blobs (Fig. 6). There were some noisy areas in the image which could lead to false representation of foreground in terms of small blobs. A sizing filter was used to eliminate the blobs of sized too small than estimated object size. Since the background is also prone to change with respect to visual appearance it was updated dynamically using the method mentioned in Eq.(2).

$$B_{t+1} = \alpha I_t + (1-\alpha)B_t \tag{2}$$

This first order running average adaptive filter [xiv] used for background estimation with value of

$\alpha=0.05$ as the learning rate. Every moving object was marked by a bounding box defining the boundary of the object. These bounding boxes were then used for feature extraction in subsequent steps.

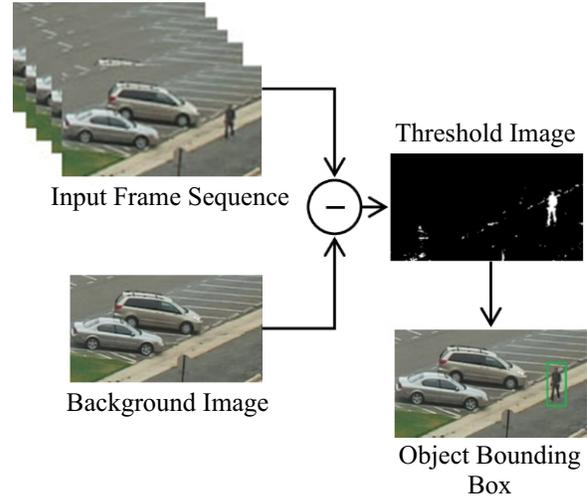


Fig. 6. Foreground extraction and area of interest localization

B. Feature Extraction

In order to have a clear human form we used a slight modification of the feature descriptor presented by [xv]. Combining edge orientation histograms and SIFT techniques into more robust descriptor as [vi] propose. We evaluated the object appearance and shape using local intensity gradient distributions, (Fig. 7). To represent the scene using histogram of oriented gradients (HOG) we resized the original bounding box to 128x64 pixels size as shown in Fig. 7. Then resized image was divided into 128 cells each having 8x8 pixels. A histogram of gradient direction was computed for each cell. With a gradient direction of -90o to 90o each histogram was quantified into 9 bins. To reduce the effect of illumination variation on the image we carried out the local contrast normalization.

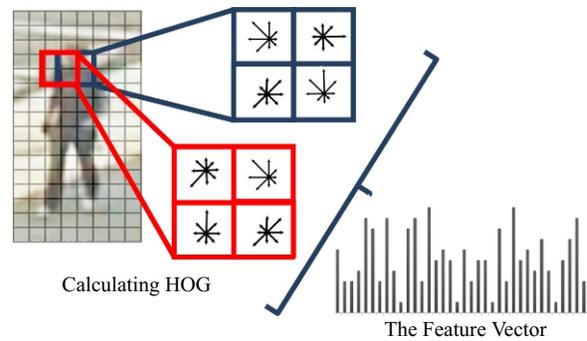


Fig. 7. Histogram of Oriented Gradients and Feature Vector Extraction

Local contrast normalization was achieved by

grouping the cells into overlapping blocks. This left us with a single block of 2x2 cells and any two neighboring blocks have 2 cells in common due to overlapping arrangement. A feature vector 36 elements was constructed using histogram of each block. The content is represented by a feature vector of dimensions 36x105.

C. Vehicle vs Human Classification

Using the feature vectors obtained in the previous segment an SVM was employed to determine whether the moving object was human or a vehicle. SVM classification was chosen because of its properties to generalize well even in higher dimensional space and it required less training samples. We used to go through the SVMs for a two-class classification problem on similar lines as mentioned in [xvi]. Given a training set $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ where feature vector $x_i \in R^d$ and label $y_i \in \{1, -1\}$ the goal of SVM was to construct a hyperplane to maximize the margin while minimize a the misclassification error. The optimal separating hyperplane $w^* \cdot x + b^* = 0$ was found under the following constraints as given in Eq.(3):

$$\min_{w, b, \xi} \frac{1}{2} w \cdot w + C \sum_{i=1}^n \xi_i \quad (3)$$

Subject to $y_i(w \cdot x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$, $i = 1, \dots, n$, $\xi_i \geq 0$, $i = 1, \dots, n$. where C refers to the penalty parameter that was used to control the tradeoff between the margin and the misclassification errors $\xi = (\xi_1, \xi_2, \dots, \xi_n)$. This became a quadratic programming problem which was solved by Lagrange multipliers. Given the decision function $f(x) = w \cdot x + b$ the Eq.(4) gave the posterior class probability as given below:

$$P(y = 1|x) \approx \frac{1}{1 + \exp(Af(x) + B)} \quad (4)$$

Most appropriate setting of parameters (A,B) was obtained from the training data as reported by [xvii]. Once the humans and vehicles were segmented out in a video frame activity recognition methods were used to identify the ongoing activities as discussed in following section.

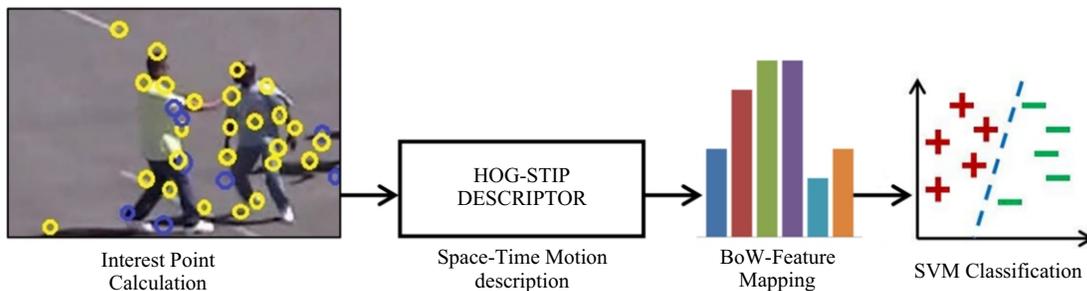


Fig. 8. Interest point calculation and activity classification using SVM

D. Parking Lot Activity recognition

In order to recognize human activity in the parking area we used the similar approach as one mentioned in [xviii]. We calculated a bag of features representation from previous section. The spatio-temporal interest point descriptor STIP [xix] was used for bag of words (BoW) collection (Fig. 8). This Bow representation was used in two class SVM in order to classify an activity. As given in Eq.(5) two-class SVM, the decision function for a feature vector \mathbf{x} of a test video had the following form as given by Eq. 5.

$$f(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) - b \quad (5)$$

$K(\mathbf{x}_i, \mathbf{x})$ was the output of the Gaussian kernel function for the features of i th training video \mathbf{x}_i and its test sample \mathbf{x} ; y_i was the event class label of \mathbf{x}_i ; α_i was the learned weight of the training sample \mathbf{x}_i ; and b was a threshold parameter. For the feature descriptors presented in (BoW) manner, it has been learned that χ^2 Gaussian kernel is more suitable defined as below

$$K(\mathbf{x}, \mathbf{y}) = e^{-\rho d_{\chi^2}(\mathbf{x}, \mathbf{y})} \quad (6)$$

$$d_{\chi^2}(\mathbf{x}, \mathbf{y}) = \sum_j \frac{(x_j - y_j)^2}{(x_j + y_j)} \quad (7)$$

In Eq.(7) $d_{\chi^2}(\mathbf{x}, \mathbf{y})$ was the distance between samples \mathbf{x} and \mathbf{y} . The performance of SVM classification was subject to a few parameters, ρ is the most important of them found in the kernel function. Data distribution itself gave a hint for proper parameter selection. Cross-validation mechanism was used to evaluate the range of parameters to choose the best one

IV. RESULTS

To test and validate our proposed method for activity recognition, experiments were performed on VIRAT benchmark dataset [xx]. VIRAT is used for large scale event detection: In our experiments we used the Release 1.0 of the dataset which was publish in the CVPR'11 activity recognition challenge as well. It has 128 videos released as test videos where total 6

different scenes are present, three same scenes like in training set with three additional scenes Fig. 3. These videos were captured at 1080px720p by fixed mount high definition camera. Size of human shape found in this data set ranged 20-220 pixels; this made 222% of the heights of video screen with average being about 9%. Results of human detection algorithm are given in Table I, along with performance comparison of

proposed human detection with state of the art methods [xxi] and [xxii] is given in Table II. The proposed activity recognition model was trained and tested for seven different types of human activities on VIRAT 1.0 data set videos. Later on these trained models were tested using parking lot videos from smart camera as well. The results of activity recognition process are given in Table III.

TABLE I
RESULTS OF PROPOSED HUMAN DETECTION ALGORITHM IN DIFFERENT TEST VIDEOS

Video Description	Number of Humans Present	Humans detected	Missed Detections	False Detection
1. Indoor parking video	130	123	7	12
2. Outdoor parking	210	199	11	19
3. Corridor	160	151	9	14
4. Bus Stop video	250	239	11	8
5. Street surveillance video	300	288	19	18

TABLE II
PERFORMANCE COMPARISON OF HUMAN DETECTION METHODS ON PARKING LOT VIDEOS

Video Content Description	Proposed Method	(Montabone and Soto, [xxi])	(Khan and Saeed, [xxii])
1. People walking by	82.5%	80.6%	77.8%
2. Group of people standing	80.1%	79.0%	75.5%
3. Person entering vehicle	79.3%	77.2%	76.8%
4. Person leaving vehicle	81.5%	78.7%	77.9%

TABLE III
RESULTS OF ACTIVITY RECOGNITION ALGORITHM ON PARKING LOT VIDEOS

Activity Description	Activities Present	Activities Detected	% Accuracy
1. Vehicle entering parking area	80	73	91.3
2. Vehicle leaving parking area	80	72	90.0
3. Person entering vehicle	75	65	86.7
4. Person leaving vehicle	75	67	89.3
5. Person loading into vehicle	33	25	75.8
6. Person opening trunk	26	19	73.1
7. Person-person interaction in parking area	45	37	82.2

A user opinion survey was conducted to gauge the usefulness of the proposed solution. A person participating in this survey had to own a car and a smart phone with internet connectivity with our application installed. A group of parking lot users comprising 250 persons from different age groups took part in this survey. After a period of one month of usage of our parking lot management there were over 6000 interactions with the smartphone application. At completion of one month of usage the participant were asked a variety of questions regarding their experience with the system. The survey questionnaire comprised of questions about user interface, ease of access and utility of the proposed system. The outcome of user

survey is shown in Fig.9.

V. CONCLUSION

Based on test results and user opinion survey findings it is concluded that video analytics and mobile phone based parking lot management system offered reliable parking solution. This framework is capable of providing seamless automated parking facility which is appreciated by vehicle owners as well as parking security staff. The proposed system automated the existing parking management system by keeping all the in/out information and parking status of vehicles. In the meanwhile it helped to facilitate the parking

security department by automatically identifying the human activities in the area. This system also saves a considerable number of vehicle miles inside parking lot area by efficiently allocating the parking place for incoming vehicle thus eventually contribute to lower fuel consumption. User behavior identification for parking patterns in large scale parking lot videos would be an interesting future extension of this work.

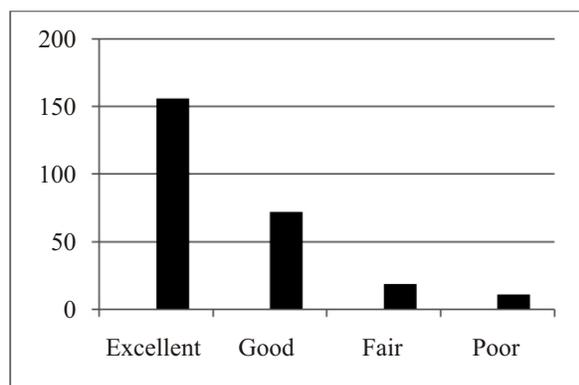


Fig. 9. Results of user survey on utility of the proposed system

VI. ACKNOWLEDGMENTS

This work has been fully supported by the Directorate of Advanced Study and Research (ASR&TD) University of Engineering & Technology Taxila.

REFERENCES

[i] W. Wang, Y. Song, J. Zhang, and H. Deng. "Automatic parking of vehicles: A review of literatures." *International Journal of Automotive Technology* 15, no. 6:(2014) pp.967-978.

[ii] L. Y. Mimbela, and L. A. Klein. "A summary of vehicle detection and surveillance technologies used in intelligent transportation systems." *Technical Report*, New Mexico State University, New Mexico.(2007)

[iii] Y. G. Jiang, S. Bhattacharya, S. F. Chang, and M. Shah. "High level event recognition in unconstrained videos." *International Journal of Multimedia Information Retrieval* 2, no. 2: (2013)pp.73-101.

[iv] S. Ali, A. Basharat, and M. Shah. "Chaotic invariants for human action recognition." *IEEE 11th International Conference on Computer Vision ICCV'07*.(2007) pp. 1-8.

[v] Y. G. Jiang, Q. Dai, X. Xue, W. Liu, and C. W. Ngo. "Trajectory based modeling of human actions with motion reference points." *Proceedings of European conference in computer vision*. (2012).

[vi] N. Dalal, and B. Triggs. "Histograms of oriented gradients for human detection." *IEEE Conference on Computer Vision and Pattern Recognition*.(2005) pp. 886-893.

[vii] B. Chakraborty, M. B. Holte, T. B. Moeslund., and J. Gonzáles. "A selective spatiotemporal interest point detector for human action recognition in complex scenes." *Proceedings of the International Conference on Computer*.(2011).

[viii] M. F. Bulbul, Y. Jiang, and J. Ma. "An Enhanced Histogram of Oriented Gradients for Pedestrian Detection." *IEEE International Conference on Multimedia Big Data (BigMM)*. Beijing, (2015) pp.389-394.

[ix] I. Laptev, and T. Lindeberg. "Space-time Interest Points." *IEEE International Conference on Computer Vision (ICCV 03)*.(2003).

[x] J. K. Aggarwal, and M. S. Ryoo. "Human Activity Analysis: A Review." *ACM Computing Surveys*, 43, no. 3: (2009) pp.16-63.

[xi] Y. A. Ivanov, and A. F. Bobick. "Recognition of visual activities and interactions by stochastic parsing." *IEEE Trans Pattern Anal Mach Intelligence*.,vol22, no. 8: (2000) pp.852-872.

[xii] M. S. Ryoo, and J. K. Aggarwal. "Recognition of composite human activities through context-free grammar based representation." *Proceedings of IEEE conference on computer vision and pattern recognition*.(2006).

[xiii] S. W. Joo, and R. Chellappa. "Attribute grammar-based event recognition and anomaly detection." *Proceedings of IEEE conference on computer vision and pattern recognition, Workshop*.(2006).

[xiv] J. Haikkila, and O. Silven. "A real-time system for monitoring of cyclists and pedestrians." In *Proceedings of the Second IEEE Workshop on Visual Surveillance*.(1999). Pp.74-81.

[xv] D. G. Lowe "Distinctive Image Features from Scale invariant Keypoints." *IJCV* 60, no. 2: (2004). Pp.91-110.

[xvi] V. N. Vapnik. *The Nature of Statistical Learning Theory*. 2nd. New York: Springer-Verlag New York, Inc.,(2010).

[xvii] J. Platt. "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods." *Advances in Large Margin Classifiers*,:(2000). Pp.61-74.

[xviii] S. Althloothi, M., H. Mahoor, X. Zhang, and R., M. Voyles. "Human activity recognition using multi-features and multiple kernel learning." *Pattern Recogn.*, Elsevier Science Incvol 42, no.5 (2014) pp. 1800-1812

[xix] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference* pp. 124.1-124.11. BMVA Press.

- [xx] S. Oh, et al. "A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video." IEEE Computer Vision and Pattern Recognition (CVPR).(2011). Image and Vision Computing, vol. 28,.no. 3, (2010)pp.391-402.
- [xxi] S. Montabone, and A. Soto. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. [xxii] M. U. G. Khan, and A. Saeed. HUMAN DETECTION IN VIDEOS. Journal of Theoretical & Applied Information Technology, vol.5,.no.2(2009).

Authorship and Contribution Declaration			
	Author-s Full Name	Contribution to Paper	
1	Engr. M. Rizwan (Main/principal Author)	Basic study Design, Data Collection, statistical analysis and interpretation of results etc. Methodology, manuscript writing, Literature review & Referencing	
2	Dr. Hafiz Adnan Habib Bravo (2nd Author)	Proposed topic, Critical Review, Revision of results, methodology and quality insurer	