Research Article

Car Surveillance Video Summarization Based On Car Plate Detection

Nouria Kaream Khoorshed, Dr. Ziyad Tariq Mustafa Al-Tai

Golden.pen.n.k@gmail.com
Ziyad1964tariq@uodiyala.edu.iq
Diyala University
Diyala University

Abstract—Today, video is a common medium for sharing information. Navigating the internet to download a certain form of video, it takes a long time, a lot of bandwidth, and a lot of disk space. Since sending video over the internet is too costly, therefore video summarization has become a critical technology. Monitoring vehicles of people from a security and traffic perspective is a major issue. This monitoring depends on the identification of the license plate of vehicles. The proposed system includes training and testing stages. Training stage comprises: video preprocessing, Viola-Jones training, and Support Vector Machine (SVM) optimization. Testing stage contains: test video preprocessing, car plate (detection, cropping, resizing, and grouping detecting test car plate, feature extraction using HOG feature. The total time of local recorded videos is (19.5 minutes), (15.5 minutes) for training, and (4 minutes) for testing. This means, (79.5%) for training and (20.5%) for testing. The proposed video summarization has got maximum accuracy of (86%) by using Viola-Jones and SVM by reducing the number of original video frames from (7077) frames to (1200) frames. The accuracy of the Viola-Jones object detection process for training 700 images is (97%). The accuracy of the SVM classifier is (99.6%).

Keywords— Car's Plate Detection, Machine Learning, Viola-Jones algorithm, SVM, Video Summarization.

1. INTRODUCTION

In recent years, sudden technical advancements in video data creation and its storage is improved a lot. So, it's closer to how efficiently we handle those data with indexing and retrieval methods. As the current activity of knowledge creation and its storage is sequential, so it consumes more space for storing. As for a particular video, there'll be repeated video content that cannot be useful and hence skip those data and extract video with less repetition [1]. A video summary is defined as a stream of still or moving pictures presenting the content of a video in such how that the relevant target is given brief knowledge while the fundamental message of the first video is preserved. There are two fundamental sorts of video abstraction techniques: the first is static video summarization which is additionally called representative frames, still-image abstracts, or static storyboard that summarizing the first video with a lot of data to a little number of frames without losing the rich information. While the second is dynamic video summarization also called dynamic video skimming, video skim, moving image abstract, or moving storyboard that summarizing the first video to video as short as possible that provides a global picture of the video. Most existing video summarization techniques are keyframe based, i.e., several frames from the first videos are extracted to represent the entire video [2][3][4]. Machine learning techniques are proved to achieve success for various image (video frame) analysis processes and object tracking. Therefore, this study used Machine Learning (ML) for the training and test process.

2. Related Works

- Dipti Jadhav and Udhav Bhosle, 2017[5], suggested a methodology for video description based on the Speeded Up Robust Features (SURF). Authors also recommend an approach based on graph theory to maximize the number of keyframes based on the objective function that the graph created by the optimized video description is a simple graph with a simple walk.
- Rajat Aggarwal, Brijesh Singh Butola, in 2016[6], used a video overview technique based on collecting keyframes and the number of frames in the overview will be as few as possible. The methodology is focused primarily on clustering the whole set of video frames into an optimum number of clusters such that the corresponding cluster heads can be viewed as keyframes for the summary collection. Selecting the optimal number of mainframes can be applied to the Expectation-Maximization system where frames are clustered (used Davies-Bouldin Index (DBI)) for clustering.

- Dong-Ju Jeong et. al., in 2017[7], proposed a two-step approach where the primary step skims a video and therefore the second step performs content-aware clustering with keyframe selection. The 1st step eliminates the most redundant frames that contain only a little new information by applying the spectral clustering technique with color histogram features. Then the result, obtain a brief video that's shorter and has clearer temporal boundaries than the first. within the 2nd step, perform coarse temporal segmentation then apply refined clustering for each of the temporal segments, where each frame is represented by the sparse coding of SIFT features. The keyframe selection from each cluster is predicated on the measure of representativeness and visual quality of frames, where the representativeness is defined from the sparse coding and therefore the visual quality is that the combination of contrast, blur, and image skew measures. The matter of keyframe selection is to hunt out the frames that have both representativeness and high quality, which is formulated as an optimization problem. Experiments result on videos with different lengths show that the resulting summaries closely follow the important contents of videos.
- Sinn Susan Thomas et.al. in 2017 [8] explained how to utilize the best security camera description system. Besides that, the search time and proposing to turn content-based video retrieval issues into a content-based image retrieval concern. Thus, multiple studies on diverse data sets were carried out in this paper to test the suggested approach to smart surveillance. The query and the database matching using NN-classifier. The video was retrieved based on features such as Graph-Based Visual Saliency. This approach used Greedy Search Algorithm can be mainly divided into two parts: First, Perceptual video description where a description in line with human vision properties plays a key role in deciding the frames to be used for the review. Second, this approach fits best for photo images, where the foregrounds in the mainframes are stitched on the backdrop to achieve a single condensed picture. This approach used two parameters to measure the performance of this system: The information rate IR reflects the volume of information in the description assessing the efficiency of the condensed process.

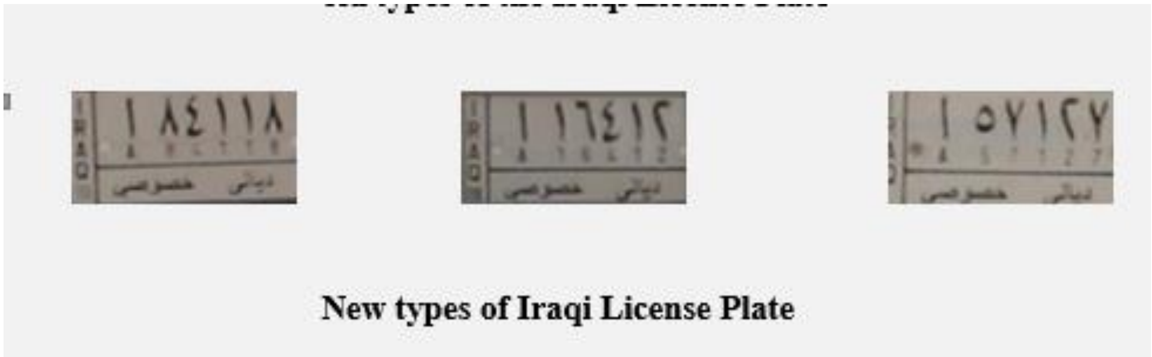
- Antti E. Ainasoja et. al. in 2018 [9], specialized in the favored keyframe-based approach for video summarization. Keyframes represent important and diverse content of an input video and a summary is generated by temporally expanding the keyframe stocky shots which are merged to an endless dynamic video summary. At the same time as this approach, keyframes are chosen from scenes that represent identically similar content for scene detection, this work proposes an easy yet effective dynamic extension of a video Bag-of-Words (BOW) method which provides over segmentation for keyframe pick. For keyframe selection, the investigate two effective approaches: local region descriptors (visual content) and optical flow descriptors (motion content). This work offered several interesting findings. 1) While scenes (visually similar content) are often effectively detected by region descriptors, optical flow (motion changes) provides better keyframes. 2) However, the acceptable parameters of the motion descriptor-based keyframe selection vary from one video to a different, and average performances remain low. To avoid more complex processing, this work presents a human-in-the-loop step where user privileged keyframes are produced by the three best methods.3) The human support and learning-free method gives superior accuracy to learning-based methods and for several videos is on a match with average human accuracy.
- Madhav Datt and Jayanta Mukhopadhyay, in 2018 [10], presented a video summarization, using convolutional neural networks (CNN) and bidirectional long short-term memory (LSTMs) to get deep features for frame representation and to model variable-range temporal sequences. Further, they introduced a parameterized loss function minimizing (Kullback-Leibler divergence) KL-divergence between the Gaussian Mixture Model (GMMs) to find out relative orders of frame importance. This work expanded extensive evaluation on a lot of benchmarks (TvSum, SumMe, and YouTube) to determine the effectiveness of this model.

3. Dataset Collection and Analysis

The dataset of this work was collected and recorded locally by the researcher in one of the private garages in (Iraq\Diyala) with different places for different vehicles. This dataset contains three videos each of which with a different long time. One of these videos was used for the testing process and the others for the training process. The angle of recording did not exceed 45o. The weather conditions for this dataset were sunny days. The type of camera that was used in this work is Canon D2000 as shown in fig. (1).

Figure(1) Canon Camera

The distance between the camera and car was not exceeding (4 m) and the angle between (0-45). The videos that were collected can be illustrated in the table (1). The type of car plate is shown in fig.(2).



Figure(2) Local Car Plate

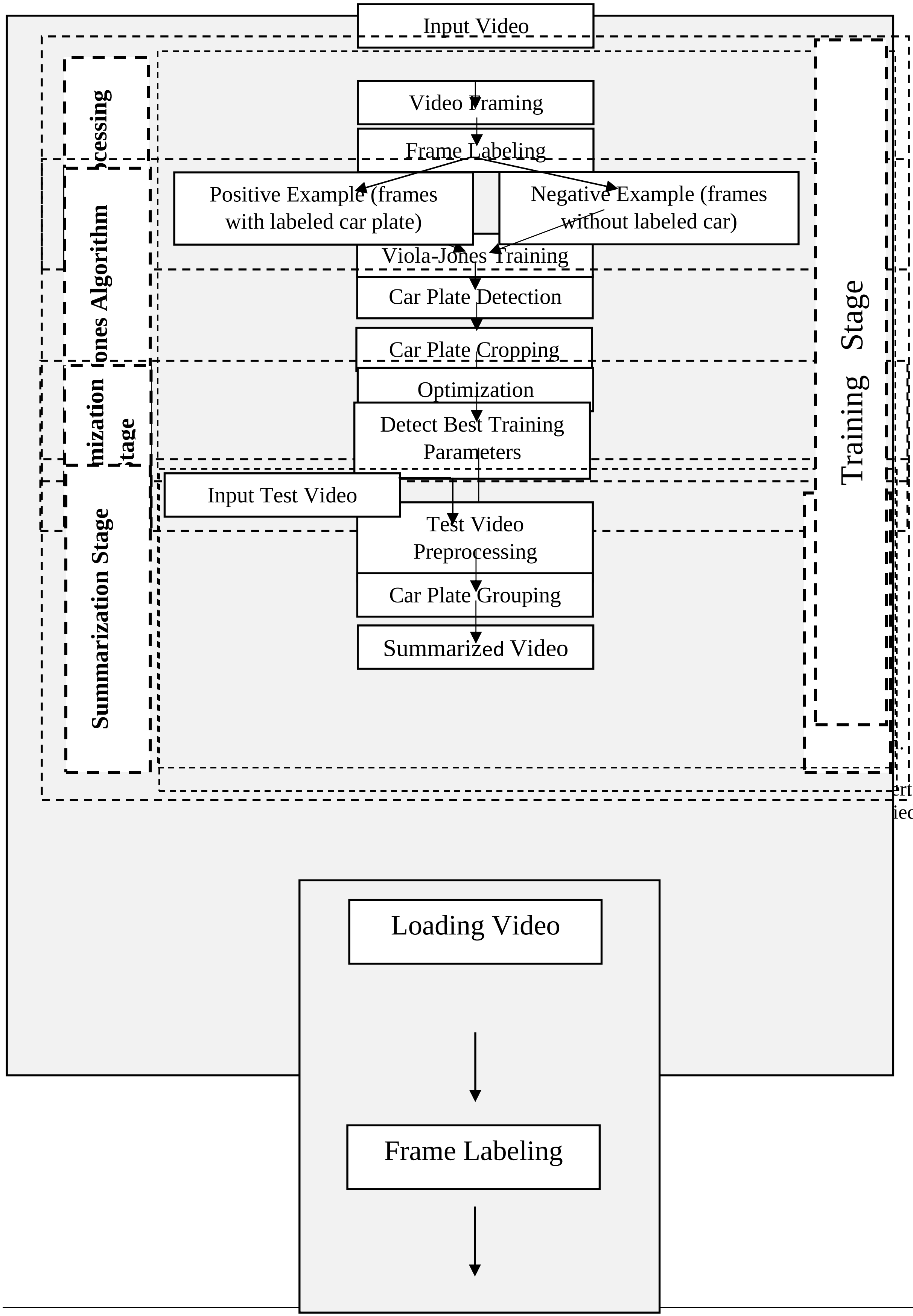
Table(1) Video of local dataset

Video	Time
Video 1	3.06 minute
Video 2	12.44 minute
Video 3	4 minute

In table (1), videos (1 and 2) are used for training, while video (3) is used for testing. In other words, The total time of recorded videos is (19.5 minutes), (15.5 minutes) for training, and (4 minutes) for testing. This means, (79.5%) for training and (20.5%) for testing.

4. The Proposed System

The proposed system is suggested to abstract a surveillance video. This summarization depends on the license plate detection. This process goes through several stages, in each stage, a group of steps is applied. In the beginning, the system is trained to obtain the best parameters through which the best results are obtained. These parameters are used and applied to the tested video. These steps can be observed in figure (3).



erts it
ied to

Figure(4) Video Preprocessing

the first step in video preprocessing is loading the recorded video by specifying its path, wherever it is stored. The second step in video preprocessing is the segmentation of the recorded video into frames. Then the recorded video is converted to a set of sequential frames as shown in formula (1).

V = [f1, f2, f3, f4, f5…………….., fN] .... (1)

Where V is the recorded video and f1.....fN represent the segmented frames. Video labeler program is used to interactively label ground truth data in a group of photographs, or sequences of pictures. Item detection and image classification scenes can be designated after rectangular regions of interest (ROIs). So, the third step of the video preprocessing is the process of detecting the car plate. This process is done by putting a label on the car plate of the extracted frame from the video. In this process, the dimensions of the plate are stored and considered as "positive frames that will be described in the next step". The frames that did not contain a car, it was left without placing any label on them so, it considered "negative frames". Then these dimensions are stored in a specific table that is retrieved in the Viola-Jones training process.

4.1.2 Viola-Jones Algorithm

Viola-Jones algorithm: When the video preprocessing is finished, the system goes to the object detection step (car plate detection). In this step, Viola-Jones algorithm is used to train the machine to find the car plate. The idea of this algorithm is shown in fig. (5).

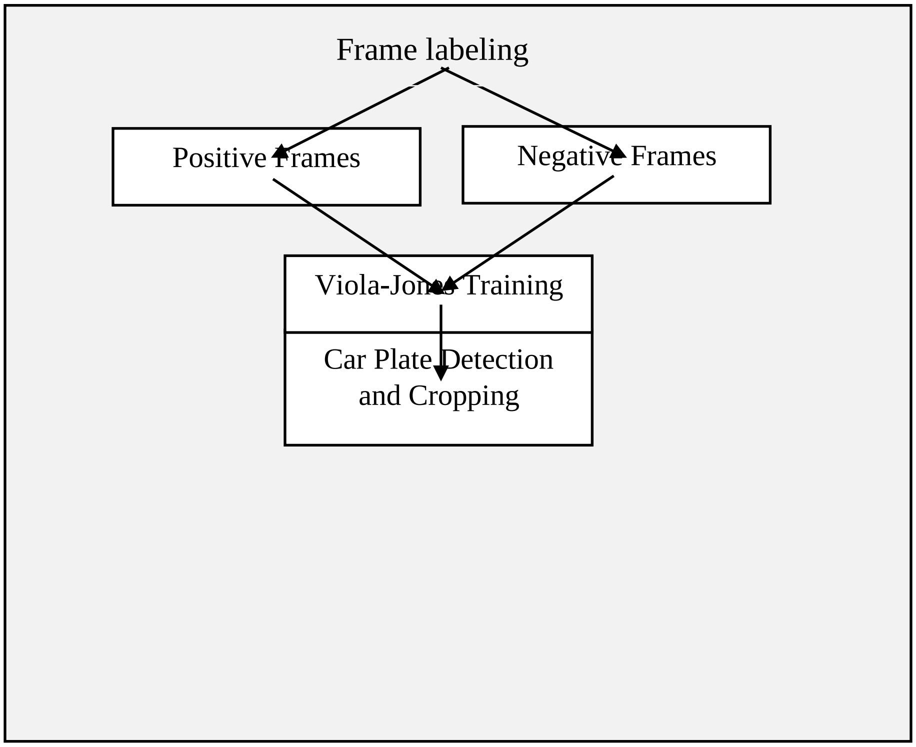


Figure (5) Viola Jones Algorithm

A. Positive Frames

Positive frames represent the frames for which a label was made, in another word it was the frames that contain a car, and a label was set for the plate of this frame. The coordinates of the car plate are loaded and stored in a table (called the "VideoLabelingTable" as a positive instance or positive example). The positive instance is a table with two columns the first one is for name of frames, while the second one is for the coordinates of the labeled plates in frames and stored as [x, y, width, height] boundaries of the box that specify object location the positive frames showed in figure (6) below.

Figure (6) Positive Frames

B. Negative Frame

These frames are the frames that do not contain label which are retrieved during the process of training. Both negative and positive frames are used with the Viola-Jones algorithm in order to learn the proposed system for differentiating between frames that contain a car plate from the ones that do not contain it fig(7) showed negative frames.

Figure(7) Negative Frames

C. Viola-Jones Algorithm Training

Both positive and negative frames (instances) from video labeling table are used in training stage. Viola-Jones with training stage uses a real-valued scalar to specify the number of negative instances which is a multiple of the number of positive instances. The proposed system is trained with different parameters such as:

Num Cascade stages: this parameter detects number of stages. Note that an increasing of the number for stages takes more to detect accurately, long time, and more training frames.

False_Alarm_Rate: this value must be greater than 0 and less than or equal to 1. When this value is decreased the result has less false detection but long training and detection time.

Features Type: This parameter gives the type of features. Note that the type of features which is used in the proposed system is HOG. The result of the training process with different parameters are stored as (.xml) file.

D. Car Plate Detection

To implement the car plate detection process, the stored dimensions must be loaded from (.xml) file which was obtained from labeling process in the pre-processing stage. Then, the original frames file that was obtained from the first step in the preprocessing stage was loaded. So, when these coordinates are applied to the frames, then the detector detects that this frame has a car and detects its plate as showed in the fig.(8) . After applying this process to all frames, the results of this step are saved. Sometimes more than one detector appears in the same frame.

Of course, not all operations of car plate detection are correct. The ratio of error depends on the value of the parameters that are applied to the frames in the training process.

Then the specified regions are cut out from the frames. These regions are with different sizes; therefore, these regions must be converted to a same fixed size.

Figure(8) Car Plate Detection

E. Car Plate Cropping

Image cropping is the process of enhancing a picture or image by eliminating the unwanted parts. After the plate of numbers was identified in step (D), the system goes to cropping step which is extracting the places that were identified by the step (D), and saving the cropped images in a specific file. The proposed system cuts identified regions regardless of whether the detection was correct or not. The size of cropped regions have different sizes, so the size of all cropped images are become with same size by using a process called resizing.

4.1.3 Optimization Step:

This step consists of several subsets as described in following subsections.

A. Support Vector Machine Algorithm

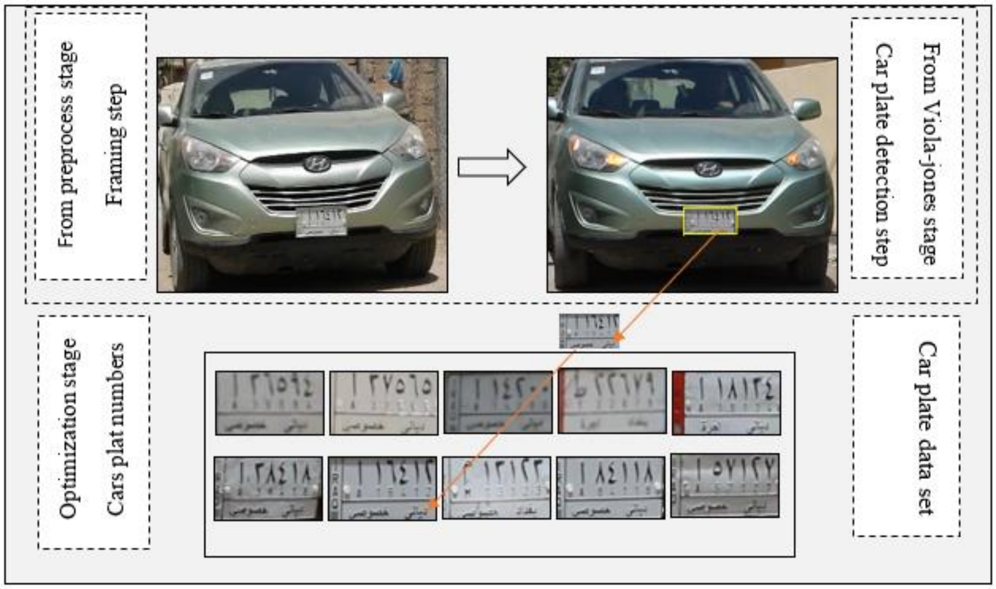
The first step in SVM algorithm is to train the proposed system by using positive and negative resized and cropped frames. This is done by loading the file that contains cropped car plate images. Two files are used to train SVM classifier, one of them contains the correct cropped images while the other contains the error cropped images. Then, features of all images are extracted using LBP[11][12]after converting these images to grayscale. SVM creates a set of feature vectors, each vector is related to a specific image. The class is assigned to value (1) if car plates are existed, otherwise the class is assigned to value (2). Using the RBF [13] as a kernel function[14].

B. Detect Best Training Parameters

Through the training process for both the Viola-Jones and the SVM algorithm the best parameters that chosen for this work were:

Viola-Jones: the best and suitable parameters that used in viola jones were False-AlarmRate which used with value 0.001 The NumCascade stage that used in this algorithm was 5 stages, number of positive frames 700, number of negative frames was 1200 and the FeaturesType was HOG[15] .

For SVM algorithm the best parameters were: Kernel_Scale after train multiple time the best result was 0.09 and Outlier_Fraction parameter with value equal to 0.01 only these parameters chosen because it gets the best result in the manner of images classification. The fig(9) below showed the optimized Car Plate Dataset.



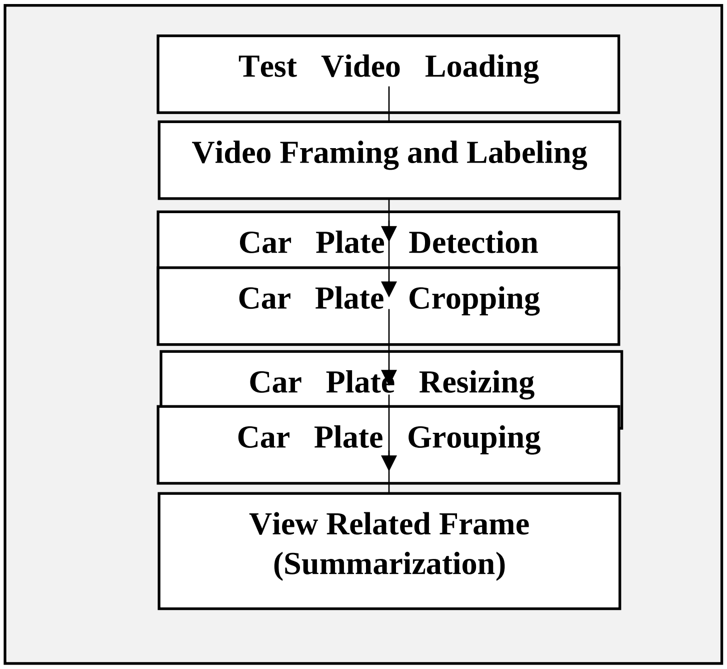
Figure(9) Optimize Cropped Car Plate

4.2 Testing Phase

Testing stage includes the achievement the objective: obtaining the summary surveillance video of all the cars that appear on the tested video.

4.2.1 Obtaining Summarization Video

The block diagram of the video summarization stage is shown in figure (10).

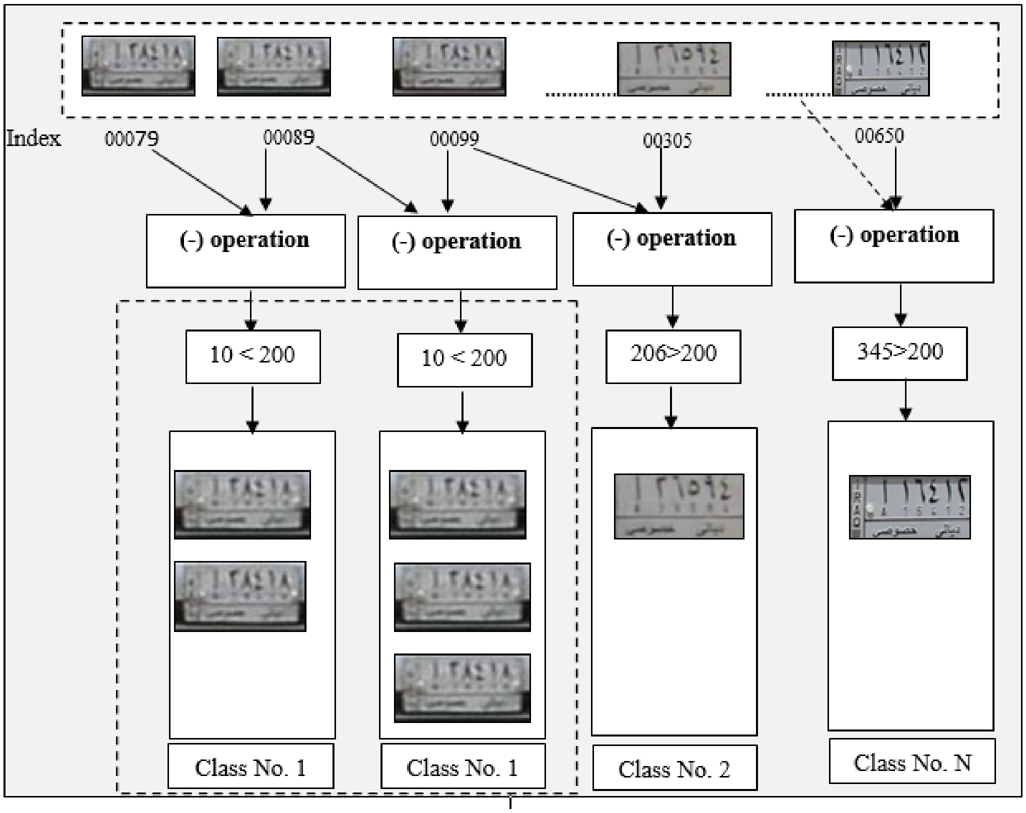


Figure(10) Video Summarization Stage

In this figure, framing process and labeling were explained as described in section (**4.1.1**). Car plate image detection and cropping processes are done using the Viola-Jones algorithm that was described in subsection(**4.1.2.D** and **E**), and the other steps are described in following subsections. As described previously in section (**4.1.2.E**) the output was a cropped image these cropping goes through the optimization step that was described in subsection (**4.1.3.A**) using the SVM algorithm. So, the final shape of these images become as images that have car plates only with the same size.

A. Car Plate Grouping

The inputs of this step are the cropped images after resizing and optimizing them. Its output is a set of groups as many as the number of cars observed in the video. The grouping process was done as shown in figure (11) dependent on the gaps between each car plate index and the next, specified threshold that separate each cropped car plate from another. The length of each group different from one to another depends on the number of detections implemented on this car plate, for example, one car plate in the plate detection and cropping process has 100 detections and cropping while another car has only 50 detections and cropping related to it. The original sequence of each frame has been preserved even through the cropping process and the path preserved. This step was very important for the summarization. Note: The purpose of grouping to check whether all plate have been discovered, Because the number of groups must be equal to the number of cars shown in the original video).



Figure(11) Car Plate Grouping

B. Summarization Step

Summarization step is taken place after grouping step, by putting different groups in one group. This is done by reading first index for first group, and when the number of images in this group ends, the images in of the next group are added and so on. This process is continued for all groups to create only one group. The original frames (that are related to each cropped image or car plate) are retrieved and displayed one after another to create a new video with required events in AVI format.

5. Experimental Result

For both object detection and enhancement processes the results were shown in this section from this paper.

5.1 Viola Jones Train Result

the system train by 23574 frames obtains from videos that described in section (3), from these frames from 400-700 frames were used as a positive frame while 1400 frame were used as negative frames, so the result of training was as showing in table(2):

Table(2) Viola Jones Training

Iter.	No. Frames	FalseAlarm Rate	No. Cascade	Time in Sec.	Error Detection	Error Cropping %	Accuracy %
1	400	0.05	4	2722.53246	112	33.9 %	66.1%
2	400	0.05	5	7880.16480	6	3.5 %	96.5%
3	400	0.06	3	1804.57829	1989	88.6 %	11.4%
4	400	0.06	4	1991.68651	398	61.1 %	38.9%
5	400	0.06	5	7172.66137	25	13.5%	86.5%
6	400	0.06	6	1536.84711	12	4.9%	95.1%
7	500	0.05	3	1082.24712	1992	90%	10%
8	500	0.05	4	2688.87618	712	78.9%	21.1%
9	500	0.05	5	7012.87950	18	7%	93%
10	500	0.05	6	8439.22400	31	11%	89%
11	600	0.06	3	964.786483	2251	96%	4%
12	600	0.06	4	2098.93290	323	44.6%	55.4%
13	600	0.06	5	7308.44572	50	25%	75%
14	600	0.06	6	8442.89194	20	81%	19%
15	700	0.001	5	2687.38299	11	3%	97%
16	700	0.05	3	1565.66811	515	50.9%	49.1%
17	700	0.05	4	5034.84884	86	20.1%	79.9%
18	700	0.06	3	1018.14443	2232	96.4%	3.6%
19	700	0.06	4	5246.68019	96	21.4%	78.6%
20	700	0.06	5	9409.75310	16	6.5%	93.5%
21	700	0.06	6	9621.40243	33	9.8%	90.2%
22	700	0.06	7	9856.53297	14	4.7%	95.3%

5.2 Support Vector Machin Train Result

For SVM training process the system used 100 frames as a car plate were 50 were not. with SVM the system used Local Binary Pattern features, Kernel_Scale, and Outlier_Fraction parameters for training and the result was as showed in the table(3) below.

Table(3) SVM Best Results

Iteration	Kernel_Scale	Outlier_Fraction	From-To	Bais	Alpha	accuracy
1	0.1	0.5	280-344	-0.3046	75*1	98.9%
2	0.9	0.9	264-344	-0.0091	78*1	93%
3	0.09	0.01	282-344	-0.3347	78*1	99.6%
4	0.18	0.015	278-344	-0.1284	54*1	98%
5	0.28	0.025	255-344	-0.0792	45*1	90%

The field (From-To) represent a test images were a tested file contain 344 images these images were mixed car plate and not from these 344 images only 283 images were a car plate were the other 61 were not car plate the table(3) showed using different value for the parameters to find best result. The best result obtained in the iteration 3.

2.3.3 Summarization Result

In this work, test video with (4) minutes long was used. The results of the framing process are (7077) frames. The results of car plate cropping after using the viola jones algorithm was (344 images) using the efficient parameters. The cropped images come in different sizes, in which they converted to fixed-size with dimensions (300*600) as shown in the figure (12).

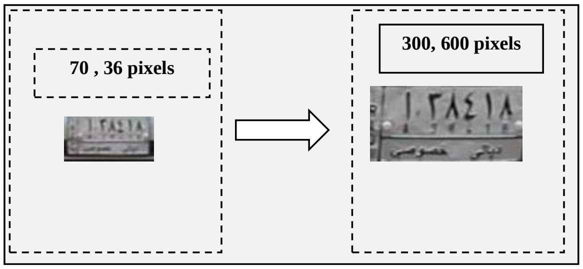


Figure (12) Cropped Image Resized

Results of Optimizing Cropped Car Plate Images:

From is (344) cropped images, (283) are correct cropped car plate images, while (61) are false cropped images, this process is done using the SVM algorithm in testing mode with the best parameter as described in table (3) (iteration 3).

Results of Car Plate Image Grouping the result of this process was (5) groups (because the tested video that used in this work contain only 5 cars) each group represent a specific car and each group has a different number of car plate images. The grouping process is shown in figure (13).

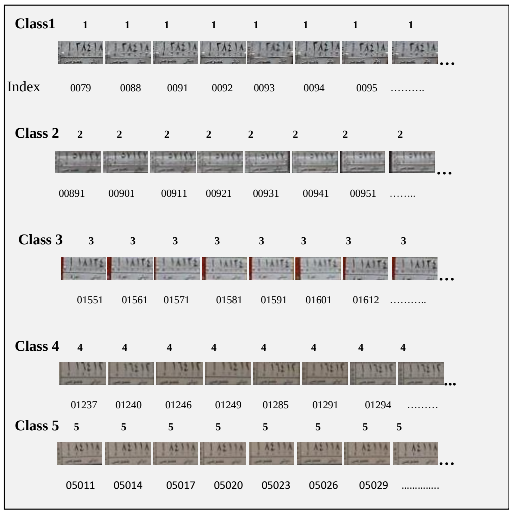


Figure (13). Car Plate Groups

The result of the summarization step was a video with 35 seconds and 1260 frames while the original test video was with 4-minutes or 240 seconds and 7077 frames. Figure (14) shows images of cars from summarized video frames.

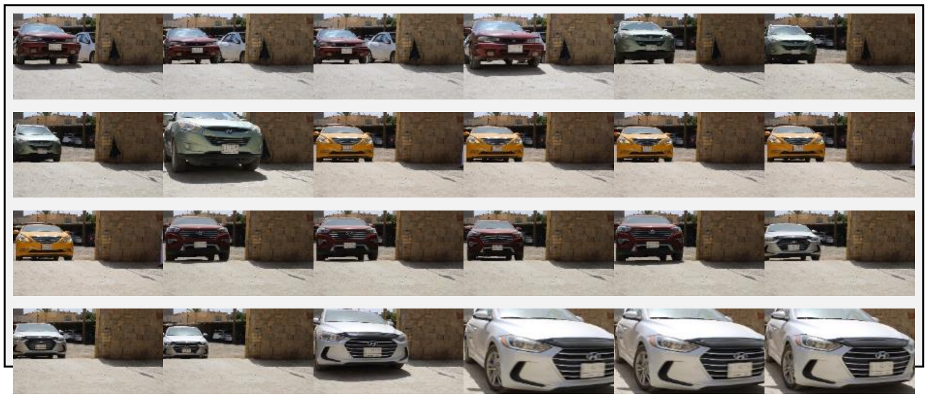


Figure (14) Summarization Result

2.4 Comparison to Other Related Works

There is a lot of research that has been addressed and is still dealing with the issue of summarizing the video using various techniques, methods, different properties, and various datasets in different fields (surveillance, sport, YouTube, .. Etc.). This work was compared with the related works. This work was the only one in the field of video summarization based on car plate detection. Where a real video of cars crossing the garage was used. Table (4) shows this comparison.

Table(4) Comparison

Authors and Reference	The method Used	Result(Summarization)
Dipti Jadhav and Udhav Bhosle, 2017.	SURF, Graph theory	85%
Rajat Aggarwal, Brijesh Singh Butola, in 2016.	Davies-Bouldin Index (DBI) for clustering	61%
Dong-Ju Jeong et. al., in 2017.	Spectral clustering, color histogram features, SIFT	76.6%
Sinn Susan Thomas et.al. in 2017 .	NN-Classifier, Greedy search algorithm	71%
Antti E. Ainasoja et. al. in 2018	Bag-of-Words	59-66%
Madhav Datt and Jayanta Mukhopadhyay, in 2018	CNN, LSTMs, KL-divergence, GMMs	43.3-60.5%
The proposed method	Viola-Jones, SVM	86%

2.5. CONCLUSION

From the results which were carried out previously, the following points are inferred:

- The number of cascaded stages (Num Cascade Stage) is an important factor in the proposed model, specifically in the Viola-Jones training table (4.6). In this table the accuracy fluctuates as the value of Num Cascade Stage was increased the accuracy of detection was also increased, but the time taken for car plate detection was also be increased.
- In the proposed model, efficient feature detection is used, because the HOG feature is utilized with the Viola-Jones algorithm. Note that Haar features are applied with most Viola-Jones algorithms. HOG feature utilization made the identification of car plate was more accurate.
- Training of Viola-Jones and SVM algorithms was used more than one time because not all cars were detected in the first training.

References

1. Srinivas, M., Pai, M. M., & Pai, R. M. (2016). An improved algorithm for video summarization– a rank-based approach. Procedia Computer Science, 89, 812-819.
2. Potapov, D. (2015). Supervised Learning Approaches for Automatic Structuring of Videos (Doctoral dissertation).
3. Furini, M., & Ghini, V. (2006, January). An audio-video summarization scheme based on audio and video analysis. In IEEE CCNC.
4. Song, X., Sun, L., Lei, J., Tao, D., Yuan, G., & Song, M. (2016). Event-based large scale surveillance video summarization. Neurocomputing, 187, 66-74.Viola, P., & Jones, M. (2001). Robust real-time object detection. International journal of computer vision, 4(34-47), 4.
5. Jadhav, D., & Bhosle, U. (2017, April). SURF based Video Summarization and its Optimization. In 2017 International Conference on Communication and Signal Processing (ICCSP) (pp. 1252-1257). IEEE.
6. Aggarwal, R., & Butola, B. S. (2016, April). Event summarization in videos. In 2016 International Conference on Computing, Communication and Automation (ICCCA) (pp. 1150-1154). IEEE.
7. Jeong, D. J., Yoo, H. J., & Cho, N. I. (2016). ***A static video summarization method based on the sparse coding of features and representativeness of frames***. *EURASIP Journal on Image and Video Processing*, *2017*(1), 1.
8. Thomas, S. S., Gupta, S., & Subramanian, V. K. (2017, July). ***Smart surveillance based on video summarization***. In *2017 IEEE Region 10 Symposium (TENSYMP)* (pp. 1-5). IEEE.
9. Ainasoja, A. E., Hietanen, A., Lankinen, J., & Kämäräinen, J. K. (2018). ***Keyframe-based Video Summarization with Human in the Loop.*** In *VISIGRAPP (4: VISAPP)* (pp. 287-296).
10. Datt, M., & Mukhopadhyay, J. (2018, October). ***Content based video summarization: Finding interesting temporal sequences of frames***. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 1268-1272). IEEE.
11. Karis, M. S., Razif, N. R. A., Ali, N. M., Rosli, M. A., Aras, M. S. M., & Ghazaly, M. M. (2016, March). Local Binary Pattern (LBP) with application to variant object detection: A survey and method. In 2016 IEEE 12th International Colloquium on Signal Processing & Its Applications (CSPA) (pp. 221-226). IEEE.
12. Heikkilä, M., Pietikäinen, M., & Schmid, C. (2009). Description of interest regions with local binary patterns. Pattern recognition, 42(3), 425-436.
13. Liu, Z., Lv, X., Liu, K., & Shi, S. (2010, March). Study on SVM compared with the other text classification methods. In 2010 Second international workshop on education technology and computer science (Vol. 1, pp. 219-222). IEEE.
14. Saeed N.A., Al-Ta'i Z.T.M. (2020) Heart Disease Prediction System Using Optimization Techniques. In: Al-Bakry A. et al. (eds) New Trends in Information and Communications Technology Applications. NTICT 2020. Communications in Computer and Information Science, vol 1183. Springer, Cham. https://doi.org/10.1007/978-3-030-55340-1_12
15. Jadhav, D., & Bhosle, U. (2017, April). SURF based Video Summarization and its Optimization. In 2017 International Conference on Communication and Signal Processing (ICCSP) (pp. 1252-1257).