EFFICIENT SEMANTIC-BASED VEHICLE RETRIEVAL IN LONG-TERM CAR PARK VIDEOS

Clarence Weihan Cheong, Ryan Woei-Sheng Lim, John See, Lai-Kuan Wong, Ian K.T. Tan*

Faculty of Computing and Informatics, Multimedia University, Malaysia (*Corresponding author)

clarence_han@hotmail.com, ryanlim0616@gmail.com, {johnsee, lkwong, ian}@mmu.edu.my

ABSTRACT

The proliferation of video data has resulted in huge potential in Big Data technology and applications for video surveillance, security, multimedia tools and analytics. However, large-scale video data over a long period of time necessitates an efficient representation such that the task of retrieving video shots is rapid and accurate. This paper proposes an efficient and comprehensive framework for semantic based vehicle retrieval from long-term car park videos. Colour and motion semantics are respectively retrieved using intuitive colour term and sketch-based trajectory querying. The contribution of this work is twofold. First, we present a strategy for extracting the dominant colour for similarity matching against the Munroe ground truth tuples. Secondly, our proposed framework introduces a unique sketch-based method of retrieving vehicle motions, which relies on user-drawn trajectories. Using the spatio-temporal atom representation for extraction from videos, our approach obtained reasonably good precision scores at very fast retrieval speeds based on one-month long of daytime car park videos.

***Index Terms*—** Video Retrieval, Sketch-based Query, Motion Trajectories, Colour Term Representations, Long-term Surveillance, Car Park

1. INTRODUCTION

The growth of surveillance industry along with its inexpensive implementation cost brings forth the rise to a constant stream of video data filling up database warehouses. Most of the time, these video data are stored for the purpose of backup and are often left unprocessed and unused, thus taking up additional storage space, only for it to be discarded or overwritten when not required.

The traditional approach of searching for a particular video shot from a huge collection is performed manually by sifting through hours and hours of video footage in a tedious manner. Of course, the task is a trivial one when there is only a single desired video shot with a definitive time, date, location and description given. However, this process takes an enormous effort when details are less clear or when all video shots with similarly described properties are desired.

In terms of surveillance video footage, there are several related works that has proposed different concepts in order to tackle the semantic extraction and retrieval process. In [1, 2, 3], the authors tested the proposed methods on relatively short test sets of several hundreds of frames up to 2 days of video data. Colour information for vehicle retrieval [4, 5] has been shown to be a stable attribute for vehicle retrieval. However, these works were highly hand crafted and was mainly tested on high quality images. As for vehicle motions representation, the authors in [1, 2, 6] applied quantization techniques to categorise motions into several cardinal directions. While useful for motion representation, these way of hard-coding directions indirectly makes it difficult for end users.

In this paper we present an efficient and comprehensive framework to semantic based vehicle retrieval from long term car park video surveillance. Particularly, we emphasise on the fact that *"long-term"* here indicates the length of video duration in which the object semantics are to be retrieved from; in this work, we use one month of videos. To our best knowledge, the longest length of videos processed is 500 hours, which was only for the task of trajectory counting [6]. For each input video, we extract object specific semantics such as vehicle colour and motion trajectories and quantize them into spatio-temporal cubes which facilitates quick retrieval.

The contributions of our work lie in the retrieval of two vehicle semantics (colour and motion), which are performed in contrasting fashion:

1. Vehicle colours are rank-recalled based on similarity between the vehicle average dominant colour and the ground truth colour values established by Munroe [7].
2. Vehicle motion tracks are retrieved by trajectory-based querying which matches a query track with all other tracks in the database (of non-uniform lengths) based on our proposed normalised Chamfer similarity ratio. We implemented this in the form of an interactive sketch-based system which allows users to draw trajectories to recall similar ones from the database.

To demonstrate the efficacy of our framework, we evaluated the reliability and speed of retrieval over a month's worth of extracted data.

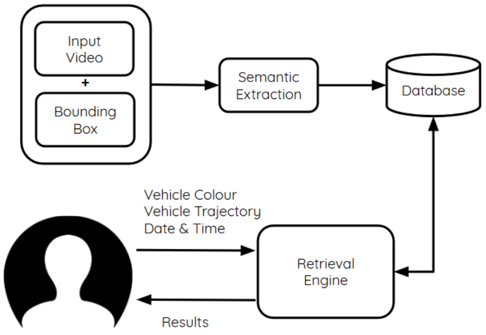


Fig. 1. Framework diagram.

2. RELATED WORKS

The extraction of vehicle colours is essential for a wide variety of applications in Intelligent Transportation System (ITS) such as crime prevention and security purposes. According to [8], colour is one of the most stable attributes of vehicles and often used as a valuable cue in some important applications. However, [4, 8] also highlighted that in a surveillance scenario, the varying illumination, complex environment factors (weather, noise) and camera viewpoint in an outdoor scene affects the classification of colours drastically. In [4], a global colour correction method is employed to suit different lighting needs in an outdoor surveillance scene. [5] took on a different approach by implementing a Homogeneity Patch Search method along with Adaboost classifier. However, these methods were only tested on high resolution data.

For retrieval systems, a number of works [3, 9] found motivation from string matching algorithms. Queries were represented in keywords such as “Red” car “turning into junction A” or “Enter” from “entrance B” to simulate natural language queries. However, [10] claims that the use of textual queries has been proven to be ineffective in video retrieval systems because keywords may not be able to capture the type of semantic content required by an end user.

In our previous work [2], we introduced a comprehensive framework for vehicle semantics extraction and retrieval for long-term car park surveillance videos. While high accuracy results were reported, the evaluation was only tested on 2 test cases, which is not sufficient to proof its reliability. Furthermore, the earlier proposed methods also showed undesirable traits as the recall rate drops significantly when the number of inputs increased.

3. FRAMEWORK

The framework in this work abides to the typical top-down approach used in ITS. In this work, the video footage and its corresponding vehicle bounding boxes are taken in as input to the system (See Section 4.1). Next, object-specific semantics are extracted from the footage and stored in the database. A sketch-based search interface was proposed to allow users to input trajectory queries as a sketch along with other semantic queries such as colour and date time information. The overview of the proposed framework is illustrated in Fig. 1.

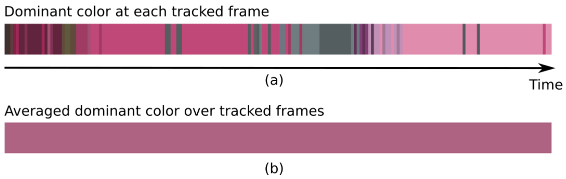


Fig. 2. (a) Dominant Colour of the Vehicle at Each Frame; (b) Average Dominant Colour (ADC)

	⬛	⬜	■	
HEX	#000000	#ffffff	#929591	
Color Term	Black	White	Gray	
	🟫	🟥	🟧	🟨
HEX	#653700	#e50000	#f97306	#ffff14
Color Term	Brown	Red	Orange	Yellow
	🟪	🟩	🟦	■
HEX	#7e1e9c	#15b01a	#0343df	#ff81c0
Colour Term	Purple	Green	Blue	Pink

Table 1. Colour Terms and the Corresponding HEX value

3.1. Semantics Extraction

3.1.1. Colour Semantic Extraction

First, the vehicle bounding boxes are shrunk by 30% as a means to crop and reduce noisy pixels at the fringe belonging to road or other vehicles. Next, a 3D HSV colour histogram is extracted from the bounding box. As colour belongs to a spectrum, the colour information is quantized into 15 Hue, 8 Saturation and 8 Value bins to form the histogram. At each frame, the histogram bin with the highest number of hits is determined to be the dominant colour. By aggregating the dominant colours (DC) obtained at each tracked frame along the trajectory, the Average DC (ADC) can be measured. This process of averaging out the DC provides robustness against illumination and pose (viewing angle) variations.

Upon obtaining the ADC of each vehicle, the next step is to determine the *colour term*. The 11 basic colour terms proposed by [11] was adopted to limit the number of colour terms. However, the definition of colour terms is insufficient as it does not provide representation of these colours in numerical terms, i.e. RGB values. We represent both the measured ADC and ground truth color terms from Munroe [7] using an RGB tuple (C_R, C_G, C_B).

The usage of 11 possible colour terms to describe an ADC

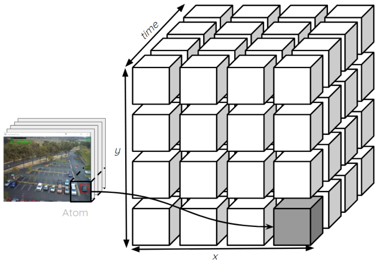


Fig. 3. Atom Based Structure. Atom's width and height were set at 32 pixels and 24 pixels respectively.

is essential as some colours bear high similarity against the others. Even though this method costs additional storage space, this implementation increases the robustness of the proposed method as compared to using typical RGB representations. A good example would be between pink and red colour; while they are essentially two different colours, these colours belong to a relatively similar hue family. Hence, the similarity score between both these colours would be rather high. This setup allows colours which are visually similar to be ranked higher than those of entirely different hues.

3.1.2. *Motion Semantic Extraction*

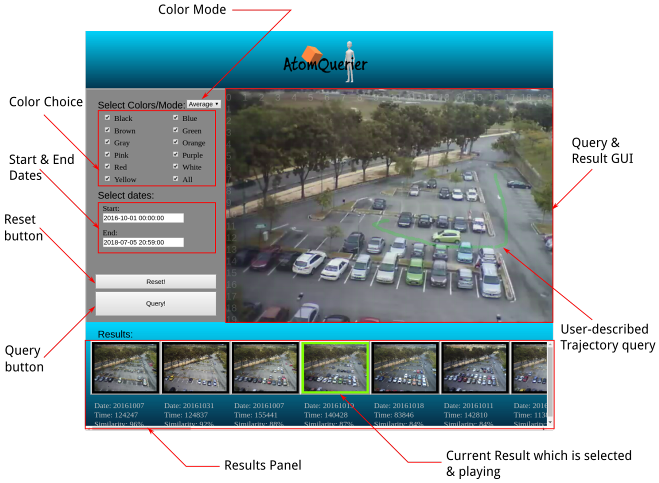
For the purpose of representing the extracted motion information, the use of spatio-temporal cubes or *atoms* [12, 2] are adopted here. As the bounding box of the vehicles is taken as the input, the motion of the vehicles can be characterised by the centroid of the bounding boxes. In this setup, each *atom* can be uniquely identified using its respective atom coordinate indices $\mathcal{A}_i(x_i, y_i, t_i)$ starting from frame i. A vehicle's trajectory is denoted as a set of tracked frame centroids,

$$P = \{\mathcal{A}_i, \mathcal{A}_{i+1}, \ldots, \mathcal{A}_{i+n}\} \tag{1}$$

where the cardinality of the trajectory $|P| = n + 1$.

In order to determine the direction of motion (of the vehicle), the sequence order based on the t-coordinates is used to obtain the motion of the trajectory, P. For each centroid location, their respective x- and y-coordinates can be inferred as the atom dimensions (width and height) are fixed and predefined. Hence, for each track, we store the sequence of the t-coordinates along with the centroids' x and y-coordinates.

As vehicles in the car park scene generally do not move in an irregular motion, the vehicle's track sequence holds sufficient clue for the reconstruction and estimation of a vehicle's trajectory within the scene. The advantage of the atom structure is that it allows a reduction in computational overhead and storage space since we only need to retrieve later using this structure instead of finer vector features.



Select Colors/Mode: Average

☑ Black ☑ Blue
☑ Brown ☑ Green
☑ Gray ☑ Orange
☑ Pink ☑ Purple
☑ Red ☑ White
☑ Yellow ☑ All

Select dates:

Start: 2016-10-01 00:00:00

End: 2018-07-05 20:59:00

Results:

Date: 20161007 Time: 124247 Similarity: 96%
Date: 20161031 Time: 124837 Similarity: 92%
Date: 20161007 Time: 155441 Similarity: 88%
Date: 20161019 Time: 140428 Similarity: 87%
Date: 20161018 Time: 83846 Similarity: 84%
Date: 20161011 Time: 142810 Similarity: 84%
Date: 2016 Time: 113 Similarity:

Fig. 4. Proposed Retrieval Engine's Search Interface

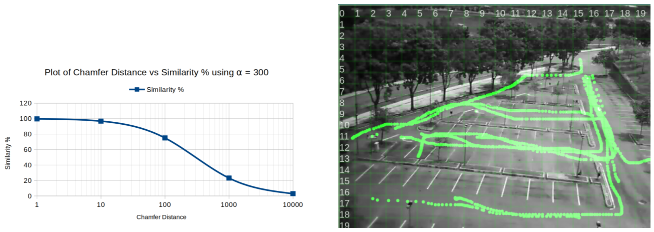
3.1.3. *Date Time Semantic Extraction*

Both date and time information are valuable information in the context of surveillance-based retrieval engines. As the input video file names contain both the time and date information, the date information can be easily extracted. Given the frame number $\mathbb{T}$, the time information for a tracked vehicle can be deduced as, $\mathbb{T}_{time} = (\mathbb{VD} \times \frac{\mathbb{T}}{\mathbb{TF}}) + \mathbb{F}_{time}$ where $\mathbb{VD}$ corresponds to the video duration of each file , $\mathbb{TF}$ is the total number of frames in the current video and $\mathbb{F}_{time}$ is the starting time information extracted from file name.

3.2. Retrieval Engine

Keyword-based retrieval engines are extremely popular in various applications. However, keywords may not be able to fully capture the query required by an end user. In this work, a combination of sketch-based and keyword-based retrieval engine is proposed. The retrieval engine takes in the following as input: Vehicle Colour, Vehicle Trajectory, and Date & Time information. The proposed retrieval engine interface is designed using a web compliant graphical user interface in JavaScript as illustrated in Fig. 4.

As the retrieval engine caters for a large dataset, the first logical approach is to minimise the search area. In this framework, the time and date semantics are first utilised to filter and narrow down the search scope. Records that fulfil the query requirements are cascaded to the next step. For a more effective retrieval, the trajectory query is given priority over the colour query. Based on the returned results, the top N results are then sorted according to the queried colour. This process allows the retrieval engine to return video shots containing the best combination of trajectory and colours. In this work, we fix $N = 200$ for ease of retrieval evaluation and checking.



(a) (b)

Fig. 5. (a) Plot of Chamfer Distance vs Similarity % using α at 300. (b) Compilation of Queries performed by users.

3.2.1. Motion Semantic Retrieval

With the output from the date and time filtering process, each individual record P can now be compared against the input query Q. First, the length of the query, $|Q|$ is compared against the length of record $|P|$. The Chamfer distance,

$$D_{Chamfer}(P,Q) = \frac{1}{\mid P \mid} \sum_{\mathcal{A}_p \in P} \min_{\mathcal{A}_q \in Q} \mid \mathcal{A}_p - \mathcal{A}_q \mid^2 \quad (2)$$

is our choice of metric to measure the dissimilarity between the input query Q and record P. Due to the asymmetry of this metric, we maximize the dissimilarity between P and Q by allowing the longer of the two trajectories (query and record) to be assigned to P and the shorter to Q, i.e. $|P| > |Q|$.

Since the difference between atom cubes term is squared in $D_{chamfer}$, the disparity between P and Q increases exponentially when they are relatively different. Hence, the obtained distance measure lies within an unbounded range [0, ∞). In order to appropriately represent the similarity score for each trajectory, this value is converted into a similarity ratio (or percentage if multiplied by 100) bounded by the range of [0, 100] by:

$$\mathcal{S} = (1 - \frac{D_{chamfer}}{D_{chamfer} + \alpha}) \quad (3)$$

where α is a free parameter empirically chosen at 300 to characterises non-linear adjustment within the range. The resulting nonlinear curve using $\alpha = 300$ can be visualised in Fig. 5(a).

3.2.2. Colour Semantic Retrieval

The distance between the track ADC and the 11 ground truth colour values of Munroe [7] is measured based on a low-cost estimation of LUV linear colour space for RGB color values introduced by [13]:

$$\Delta Colour_{LUV} = \sqrt{(2 + \frac{\bar{r}}{256}) \times \Delta R^2 + 4 \times \Delta G^2 + (2 + \frac{255 - \bar{r}}{256}) \times \Delta B^2} \quad (4)$$

where $\bar{r}$ is the average Red channel value, while the difference along each channel ΔR, ΔG and ΔB is denoted as:

$$\bar{r} = \frac{C_{1R} + C_{2R}}{2} \quad ; \quad \begin{aligned} \Delta R &= C_{1R} - C_{2R} \\ \Delta G &= C_{1G} - C_{2G} \\ \Delta B &= C_{1B} - C_{2B} \end{aligned} \quad (5)$$

For further speed-ups, the distance calculations between the ADC and the ground truth colours can be performed offline, so the retrieval system only needs to sort the results based on the distances. In the proposed system, this step is skipped if all colours are selected for the query.

4. EXPERIMENT

4.1. Dataset

The dataset used in this work is the long-term car park video dataset collected by [2]. The input videos contains 1 month of data collected from 8.30am to 6.30am (10 hours) excluding weekends (Saturdays and Sundays). The bounding box of the vehicles were obtained using adaptive background learning and frame differencing background subtraction methods proposed in [14]. The 33.4GB dataset was recorded in H.264 / MPEG-4 AVC format with a resolution of 640×480 pixels at 10 *fps*.

4.2. Experiment Methodology

The proposed methods were implemented and evaluated on an Intel i7 machine with 16GB RAM, GeForce GTX 1060 GPU. As the focal point of the proposed method revolves around the extraction of vehicle colour and trajectory semantics, both components were assessed and evaluated individually to better understand the performance, effectiveness as well as weakness.

4.3. Evaluation Metrics

As the final output of this work is an end user facing retrieval system, the evaluation process took on an empirical user study approach. This unbiased approach provided users full control of performing any kind of query, and provide feedback for each result. Six volunteers were tasked to perform queries on the retrieval system for both the colour and trajectory of the vehicles and provide relevance score REL_p for p retrieved results. As the data was only partially annotated, we are not able to compute the Recall (and thus the F_1 score) as we do not have full knowledge of all relevant video shots for any possible query.

One desirable characteristic of a good retrieval engine is to return results such that results which are more relevant are ranked higher. Users were asked to evaluate the REL on a scale of 1 to 5, if a query is relevant (3 and above) or not relevant (2 and 1). To facilitate evaluation using *Precision@K*

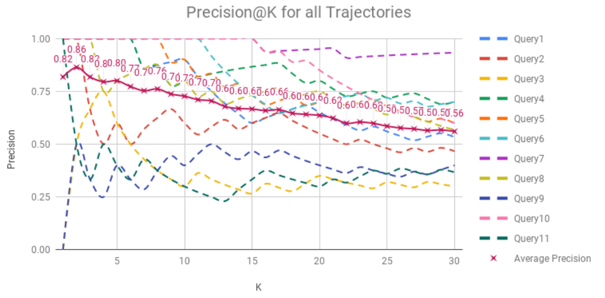


Fig. 6. Precision@K for Motion Retrieval.

metric, each relevant result is assigned 1 while the remaining ones are assigned 0. As such, given a set of relevant results R at rank position K, the *Precision@K*= $\frac{\sum^{K} R}{K}$. We also measure the performance using normalised Discounted Cumulative Gain (*nDCG*):

$$DCG_p = \sum_{p=1}^{k} \frac{REL_p}{\log_2(\max(p,2))} \qquad (6)$$

As the DCG metric produces an unbounded value ([0, ∞) range), it is not suitable for use when averaging across multiple independent queries. To achieve a bounded range of [0,1], the DCG metrics is normalised by dividing with $IDCG_P$, the DCG score at Rank P in the ideal ranking order.

5. RESULTS

The results of both retrieval tasks are reported in this section.

5.1. Motion Retrieval

First, the volunteers were given full control and were asked to draw a trajectory as an input query. Next, they were asked to rate 30 randomly ordered results. This was done to remove bias judgement on how the results should be ordered, hence increasing the impartiality and credibility of the overall results. The queries provided by the volunteers is shown in Fig. 5(b).

Fig. 6 shows that the retrieval engine is able to retrieve video shots that were quite relevant to the user's query, with a Precision of 76% at K=30. The *Precision@K* results shows a steady decline in its precision as K increases, an expected outcome in retrieval tasks.

Next, the ability of the retrieval engine to sort the results were measured. On a whole, the average *nDCG* score lies around the 83% region.

5.2. Colour Retrieval

A total of 330 video snippets (30 for each of the 11 common colours) were presented to the volunteers. These video snippets were selected using the top 30 results from the database

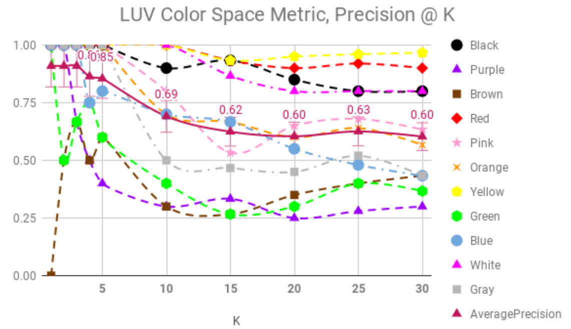
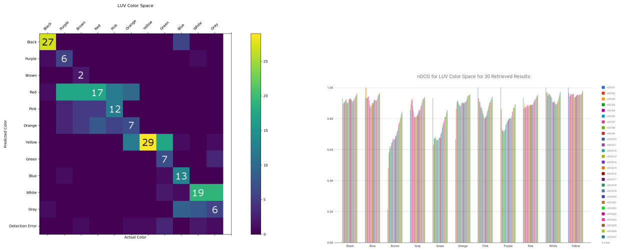


Fig. 7. Precision@K for Colour Retrieval.



(a) Confusion Matrix.

(b) nDCG.

Fig. 8. Confusion Matrix & nDCG for Colour Retrieval.

with the highest probability scores. The volunteers were then asked to determine the relevance of the matched colour term to that particular vehicle for each of the 330 snippets on the same scale of 1 to 5. At $K = 5$, the precision of the colour retrieval is reported at 85% (Fig. 7). Fig. 8(b) shows that the results consistently stays in the high 80-100% range.

5.3. Retrieval Speed

The speed of a retrieval engine in returning results is important to determine its feasibility for real-time systems. In this experiment, the time taken to perform the queries were measured, regardless of the result. A baseline (BL) measurement of the proposed method was performed with the default settings: (i) Retrieval of 30 results, (ii) Trajectory length of 10 atoms, (iii) Includes all colours, (iv) Sort first 200 colour matches. Under the baseline setting, the proposed method ignores all colour information during the retrieval process as only the trajectory query along with the time and date input affects the final results. In order to simulate the performance of the proposed method over a longer period of time, 6 months of synthetic data was generated by duplicating the existing data with different time and date information.

Table 2 summarises the retrieval speed of the proposed method using various settings (BL results in yellow row). The collection of results were done by repeating and averaging out 20 executions of queries. It can be observed that the BL settings achieved an average of 1.24s when retrieving 30 results

Table 2. Retrieval Speed with Varying Parameters

Data Size	Result Size	Traj. Length	No. of Colours	Sort Colour Size	Avg Retrieval Speed (ms)
1 month	30	10	11	200	1246.7720
	30	10	1	200	1382.6237
	30	30	11	200	1301.3777
	100	10	11	200	1282.0485
	30	10	11	500	1374.3655
	30	10	1	500	1580.7543
	30	30	1	500	1641.2180
	30	10	5	500	1972.2770
6 month	30	10	11	200	7745.1803
	30	10	1	200	7874.1965
	30	10	5	200	8314.7145
	30	30	5	200	9369.6657

Table 3. Comparison against other methods

Method	Video Duration	Video Size	DB/Index Size	Compression Ratio	Retrieval Speed (s)
DP + LSH [12]	13.8 minutes	1.16 GB	153 KB	7581	2.9
DP + LSH [12]	13.8 minutes	529 MB	65 KB	**8138.46**	4.32
Proposed method	200 hours	33.4 GB	16.3 MB	2049.07	**1.2**

given an input trajectory of 10. We observe that the length of input trajectory affects the processing speed more than the number of displayed results. Consistent with the nature of Chamfer distance metric, the proposed method would have to spend more time running through each trajectory in the set as the query length increases.

While results in Table 3 shows that our proposed method is quick, we note that direct comparisons are difficult as different datasets were used. The proposed method did not obtain high compression rates as high as [12], but searching through 200 hours of data can be accomplished in just 1.2s. Such efficiency is essential when dealing with large datasets.

6. CONCLUSION AND FUTURE WORKS

This paper presents a framework for extracting meaningful vehicle semantics from car park surveillance videos, which can then be efficiently retrieved based on color terms and trajectory sketches. Overall, the proposed framework demonstrated reasonably good performance and rapid retrieval speeds from over 200 hours of video data. As the retrieval engine was developed using web-based technology, it is easy for such tools to be implemented on modern web browsers which are OS-independent. Given that the trajectory information was stored in a pseudo-Cartesian coordinate system, the proposed method can be extended to various scenarios of similar properties.

7. ACKNOWLEDGEMENT

This work is supported in part by Telekom Malaysia Research & Development Grant No. RDTC/160903 (SHERLOCK) and Multimedia University.

8. REFERENCES

[1] Rogerio Schmidt Feris, Behjat Siddiquie, James Petterson, Yun Zhai, Ankur Datta, Lisa M Brown, and Sharath Pankanti, "Large-scale vehicle detection, indexing, and search in urban surveillance videos," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 28–42, 2012.

[2] Clarence Weihan Cheong, Ryan Woei-Sheng Lim, John See, Lai-Kuan Wong, Ian KT Tan, and Azrin Aris, "Vehicle semantics extraction and retrieval for long-term carpark video surveillance," in *Int. Conf. on Multimedia Modeling*. Springer, 2018, pp. 315–326.

[3] Bashirahamad F Momin and Tabssum M Mujawar, "Vehicle detection and attribute based search of vehicles in video surveillance system," in *Int. Conf. on Circuits, Power and Computing Technologies*, 2015, pp. 1–4.

[4] Jun-Wei Hsieh, Li-Chih Chen, Sin-Yu Chen, Duan-Yu Chen, Salah Alghyaline, and Hui-Fen Chiang, "Vehicle color classification under different lighting conditions through color correction," *IEEE Sensors Journal*, vol. 15, no. 2, pp. 971–983, 2015.

[5] Yoosoo Jeong, Kil Houm Park, and Daejin Park, "Homogeneity patch search method for efficient vehicle color classification using front-of-vehicle image," in *IEEE Int. Conf. on Imaging Systems and Techniques*, 2017, pp. 1–5.

[6] Adrien Lessard, Francois Belisle, Guillaume-Alexandre Bilodeau, and Nicolas Saunier, "The countingapp, or how to count vehicles in 500 hours of video," in *IEEE CVPR*, 2016, pp. 16–24.

[7] Randall Munroe, "Color survey results," *Online at http://blog. xkcd. com/2010/05/03/color-surveyresults*, 2010.

[8] Qiang Zhang, Jiafeng Li, Li Zhuo, Hui Zhang, and Xiaoguang Li, "Vehicle color recognition with vehicle-color saliency detection and dual-orientational dimensionality reduction of cnn deep features," *Sensing and Imaging*, vol. 18, no. 1, pp. 20, 2017.

[9] Boxiong Yang, Jing Huang, and Yuqi Yang, "Semantic description and information retrieval research of surveillance video in smart transportation system," in *Int. Conf. on Electromechanical Control Technology and Transportation*, 2015.

[10] Hrishikesh Bhaumik, Siddhartha Bhattacharyya, Mausumi Das Nath, and Susanta Chakraborty, "Hybrid soft computing approaches to content based video retrieval: A brief review," *Applied Soft Computing*, vol. 46, pp. 1008–1029, 2016.

[11] Berlin Brent and Paul Kay, *Basic color terms: Their universality and evolution*, Univ of California Press, 1991.

[12] Gregory Castañón, Mohamed Elgharib, Venkatesh Saligrama, and Pierre-Marc Jodoin, "Retrieval in long-surveillance videos using user-described motion and object attributes," *IEEE Trans. on CSVT*, vol. 26, no. 12, pp. 2313–2327, 2016.

[13] Thiadmer Riemersma, "Colour metric," *Colour metric: https://www.compuphase.com/cmetric.htm*, 2018.

[14] Ryan Woei-Sheng Lim, Clarence Weihan Cheong, John See, Ian KT Tan, Lai-Kuan Wong, and Huai-Qian Khor, "On vehicle state tracking for long-term carpark video surveillance," in *IEEE Int. Conf. on Signal and Image Processing Applications*, 2017, pp. 368–373.