Occupant monitoring system for traffic control in HOV lanes and parking lots

J. Javier Yebes, Pablo F. Alcantarilla, Luis M. Bergasa, A. González

***Abstract*— This paper presents the basics of two computer vision methods for an occupant monitoring system to be integrated as a vehicle-to-infrastructure cooperative system. The main objective is to detect passengers' faces because it is the most appealing characteristic of occupants inside a vehicle. Thus, two classification methods related to visual appearance are described and compared in this paper: Viola and Jones face detector and Bag of visual words. They both imply training and testing stages in order to build and test a complex classifier. Inside a vehicle, an onboard camera submits images through a wireless communication module to the receiver located at the infrastructure side, i.e. automated charging machines in toll highways or parking lots in the city centres. They will provide required processing resources to analyse images applying the computer vision methods presented in this paper. The obtained occupants valuable information will serve to perform different enforcement policies depending on the number of passengers.**

I. Introduction

The monitoring and control of traffic volume is becoming a constant social, economic, and environmental pressure in the industrialized countries because of current infrastructure strain under an increasingly mobile population. The viability of high-occupancy vehicle (HOV) lanes for easing traffic congestion, and hence maximising traffic flow, has been proven in countries worldwide. USA, Australia, and Canada have had HOV installations for some time, controlling the flux of traffic into their most densely populated areas. To date, all enforcement has been manual, i.e. a police officer counting the occupants in a vehicle as it passes by. Studies have concluded that manual enforcement is typically only 65% accurate. Lighting and environmental conditions, skin tone, location of the occupants and the alertness of the officer are all variables affecting the accuracy of manually collected data. An automatic vehicle occupant counting system could replace human counters and facilitate the gathering of statistical data for traffic operations management, transportation planning, and construction programming. Also, it could give law enforcement a technical means to perform the HOV lane monitoring task more effectively, as well as facilitate enforcement of a system to allow single-occupant vehicles to use the HOV lane for a fee [1]. This concept can be applied to parking lots in city centres too. Promoting the use of high occupancy vehicles by means of reduced prices and discounts, there will be less individual drivers which will lead to cut down CO2 emissions, also decreasing the pollution in city centres.

J. Javier Yebes, Pablo F. Alcantarilla, Luis M. Bergasa and A. González are with Department of Electronics, University of Alcalá. Alcalá de Henares, Madrid, Spain. e-mail: {javier.yebes; pablo.alcantarilla; bergasa;alvaro.g.arroyo}@depeca.uah.es

In this paper we present the basis for an occupant monitoring system for traffic control in HOV lanes and city centres, based in computer vision techniques using a cheap, non-intrusive VGA camera located inside vehicles. A communication module will receive the captured frames which will be wirelessly submitted to a server at the infrastructure side in order to process them and collect data about onboard occupants, e.g. the number of occupants inside the vehicle.

In the remainder of the paper, we review vehicle occupancy technologies and other related works in Section II, we show the description of our system in Section III, how the video camera is located inside the vehicle, and also the camera lenses that we use. In Section IV, we briefly review two classification methods: Viola and Jones face detector [2], and a particular Bag of visual words framework [3] for face detection. Finally, we show a comparison of both methods for our application in Section V, and remark final conclusions and future works guidelines in Section VI.

II. Related Work

Some automatic systems for occupancy detection have been under development in the last years. Most of them are based on cameras placed on the infrastructure. Thermal imaging systems have been tested but glass does not transmit waves in the range 4mm - 12mm of electromagnetic spectrum efficiently. Additionally, the application of athermic metallic coatings on new vehicles prevents transmission of heat through the windscreen. On the other hand, face recognition systems in the visible range (0.4mm - 0.7mm) suffer heavily from false positives, whereby the number of occupants is highly exaggerated. Many of the factors disrupting accurate manual detection, such as shadows and changes in light conditions, similarly fool these face recognition systems [4]. NIR range (0.4mm - 3mm) is almost completely transmitted by vehicle windscreens, even those fitted with tints or 'privacy' glass. Waves in this range are completely absorbed by human skin and completely reflected by hair, clothing, and upholstery. Face recognition systems

based on NIR range are quite good but suffer from false positives, e.g. 'dummy' passengers, pets, luggage, etc. commonly found in vehicle seats [5]. Besides, NIR cameras require expensive laser diodes to allow the system to function in a low natural light [6].

Currently, there is a commercial system [7] that overcomes the remarked limitations using a combination of infrared and visible images. The technical specifications claims that its effectiveness is high under difficult conditions. However, the camera requires enough lighting to function and the estimated cost for the system is about 100.000 pounds for one lane.

Recently the focus has shifted to in-vehicle sensing. The need for this technology has emerged out of legislation dealing with passenger airbag deployment and baby seats. The study [1] considers in-vehicle detection more suitable for enforcement. This is a mean-term solution because of the legacy of the current vehicle fleet, means a 10-15 year period for including such technologies into all new vehicles sold [5].

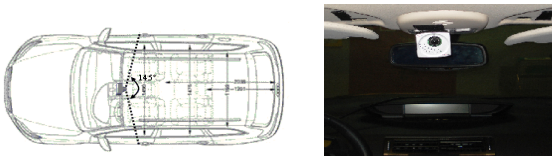
Four in-vehicle occupant detection systems are generally used today and are outlined in the synthesis report [8]. These are either a) systems sensing the force exerted on the seat rails or on a seat mat, b) systems sensing electric or magnetic fields, c) systems using optical range sensing by light time-of-flight methods, and d) optical sensing systems using cameras. Seat occupancy detection in current vehicles is exclusively realized by occupant-weight detection. Nevertheless, these systems can be deceived by heavy objects and moving persons while lifting from the seat. Recent works are focused on smart airbag deployment [9] and in-vehicle passive safety [10] in general. They are intended to detect and even classify the occupants inside a vehicle in order to contribute adaptive restraint systems in cars.

We have a great interest in this researching area and our occupant monitoring approach is intended to give valuable information about the occupants inside the vehicle.

III. System's Description

The sensing techonologies indicated in [8] are more expensive than optical sensors applied to occupant monitoring tasks, because image processing does not need further hardware upgrades once a camera has been installed inside the vehicle. Thus, our first occupant monitoring approach focuses in passengers counting, but it could be easily upgraded to more specific occupant classification using image processing algorithms. The in-vehicle system described in this section proposes the use of a wide angle lens mounted on a cheap digital camera, with low power consumption and small size. It is less costly and easier to integrate than the omnidirectional camera proposed by Wender [11]. Besides, the field of view of the lens allows to frame all the passengers in the image. Stereo cameras are not considered because we do not build a 3D model of the in-vehicle scene, hence one camera with a high field of view is enough to capture occupants' faces. The captured frames will be lately processed at infrastructure side by some classification method based on appearance features extraction and analysis, to detect the presence of passengers in the vehicle.

In addition, it is required to find a good location and orientation for the camera in order to obtain the best possible scenario, which includes all the car occupants in the images preventing from occlusions, e.g. back passengers' faces occluded by front car seats and partial faces because of proximity to image boundaries. Thus, we have checked different positions and orientations to determine the final ubication as depicted in Fig. 1. The camera is attached to the roof, some centimeters in front of the rear-view mirror, with a small inclination angle (less than 45°) to frame the passengers.



(a) (b)

Fig. 1. Camera location inside the vehicle: (a) Camera field of view (b) The camera in front of the rear-view mirror.

Considering camera and lens together, the system specifications are as follows:

1) Camera Unibrain Fire-i, VGA resolution, CCD size 1/4", typical operation power about 900mW and dimensions of 62x62x35mm. It captures grayscale images at 30 frames per second.
2) Fisheye Lens Sunex DSL215, 185° field of view (FOV).
3) Because of camera CCD size, the resulting field of view angle is aproximately 145°, which is enough for our application.

Fig. 2 is a sample frame recorded from the in-vehicle camera.

Fig. 2. Sample frame captured from the camera.

The image is distorted due to fisheye lens, but undistortion process has been proved not to be useful because of computational requirements and appearance deformations.

The occupants detection methods will evaluate the presence of a face individually for each passenger. Hence,

every frame is empirically split in regions of interest surrounding each vehicle occupant, as depicted in Fig. 3.

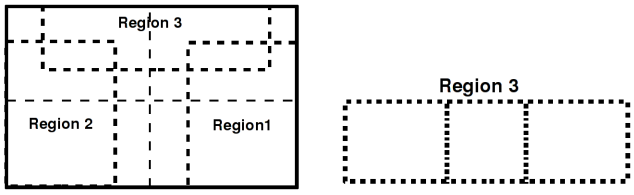


Fig. 3. Frame division in regions of interest: (a) Full frame (b) Region 3 sub-divisions.

IV. Occupants detection methods

The detection of passengers in the vehicle is based on appearance, since people faces are the most appealing characteristic of occupants inside a vehicle. This section deals with classification of every region for determining whether a face is present or not. Two methods based on visual appearance features analysis are presented: Viola and Jones face detector [2] and Bag of visual words framework proposed by Nistér [3].

A multiframe approach is proposed to process the detection results of these methods in order to obtain the final number of occupants. Detection results in consecutive frames for each image region are accumulated in a circular buffer, which filters the detected number of occupants preventing from false positives and false negatives since the number of persons in the vehicle does not change in the short term. We do not use tracking methods because faces location or moving analysis are not the goal.

A. Viola and Jones Face Detector

Viola and Jones face detector uses a rejection cascade of weak classifiers as depicted in Fig. 4, where every node defines a simple decision tree that checks different sets of attributes (called "Haar features") obtained from the image patch under analysis.

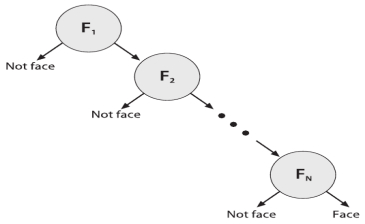


Fig. 4. Viola and Jones face detector. Rejection cascade of nodes.

An input patch will be classified as a face only if it passes tests in all the cascade nodes, whereas most of non-face patches will be rejected by the early nodes. Besides the cascade framework, there are two important additional contributions as explained in [12]: the integral image representation allows a quick computation and facilitates the comparison throughout the cascade between database features and the extracted image patch features. The learning algorithm to ensemble the cascade is based on AdaBoost [13], which sorts the nodes efficiently increasing the performance of the algorithm.

However, Viola and Jones algorithm can only reliably detect frontal and profile faces, and can not handle strong non-linearities in terms of the different possible appearances that are obtained for an occupant monitoring application. For example, under normal driving conditions changes in viewing direction, lighting, lens distortion, etc. can dramatically decrease the performance of the Viola and Jones face detector. For this reason, we try to handle the great appearance variability in faces by means of a Bag of visual words [14] framework.

B. Bag of Visual Words

This method proposes a smart recognition scheme that can scale efficiently to a large number of faces appearance. Fig. 5 depicts an example of the diversity of faces appearance under usual driving conditions.

Fig. 5. Diversity of faces appearance under normal driving conditions.

SIFT descriptors [15] representing different objects are extracted from local image regions in order to characterize face appearance. Then, these descriptors are quantized into visual words. Our approach is similar to the seminal work in image recognition presented by Nistér and Stewénius in [3]. The aim is to build a dictionary using a set of visual words extracted from training images, and later in a classification step, to search visual words from new images in that dictionary. Basically, the idea is to know if a visual word represents a face or a different thing checking its meaning in the visual dictionary or vocabulary tree of visual words. Thus, we define two categories: faces and background.

The vocabulary tree defines a hierarchical quantization that is built by means of hierarchical k-means clustering algorithm [16]. The vocabulary tree is obtained off-line in a batch process, and a large number of representative descriptor vectors are used in the unsupervised training of the tree. During run-time, the descriptor vectors for each of the new input frames are computed and then each descriptor is propagated down the tree by comparing at each level the descriptor vector to the k-candidate cluster centres and choosing the closest one. Then, using an intelligent nearest neighbors voting scheme, it is determined which database category the descriptors belong to.

The recognition quality is evaluated through retrieval from a database with ground truth data consisting of

known groups of images of the same class divided into training and testing subsets. One group is the *face category* that represents faces under different viewpoint, rotation, scale and lighting conditions. The another is the *background category* that includes empty seats under similar conditions.

V. Experimental Results

We have perfomed several tests of the two classification methods and the results are presented in the following Subsections V-A and V-B. We have built a dataset composed of 1,400 image patches corresponding to the region splitting process described in Section III. Every image patch in the dataset has different sizes according to the empirical values estimated for each region of interest around each vehicle occupant. They have been extracted from several frames of 640x480 pixel resolution. These frames belong to some recorded videos inside a vehicle, at daytime and under a variety of environment conditions.

The dataset has been divided into two sets: 700 patches where a face appears and 700 more representing background, empty seats. Fig. 6 displays a random subset of patches from each category.

(a)

(b)

Fig. 6. Sample images patches for each category: (a) Faces (b) Background.

All the images in the dataset have been manually selected in order to get several different cases:

- Illumination changes due to car movement and sun light rays penetrating through car windows, partial illuminated areas and shadows over the seats and passengers. Different weather conditions, e.g. cloudy, sunny.
- Several users with different face appearance including people with glasses, sunglasses, caps, etc.
- Changes in face direction and orientations. Natural face movements to look upside, downside, outside, etc.
- Partial face occlusions due to car seats, heads or hands movement.

Additionally, all images are radially distorted because of the fisheye lens. In fact, as we will explain in the next section this is one of the reasons why Viola and Jones face detector has trouble finding faces in the images.

A. Viola and Jones face detector

The first occupant monitoring system approach we propose is based on Viola and Jones face detector, thus we have implemented a real time application that analyses the recorded videos. A video is processed frame by frame in an offline way, using standard databases of frontal and profile faces to apply the Viola and Jones face detector for each region of interest around occupants. Then, the detection result of every region for each frame is pushed back in a circular buffer with a maximum capacity of 600 frames. In order to get the final number of occupants inside the vehicle, the data in the buffer is computed every 200 new frames. The buffer and sliding window sizes correspond to 40s and 13,3s of video at 15fps respectively. Those values have been chosen because the occupant monitoring application for car safety on road assumes a stationary number of occupants for a long period of time. Nevertheless, due to the improvements using Bag of visual words method, that condition can be relaxed as it will be shown in Section V-B. The last step of the results filtering consists on the establishment of an empirical threshold to determine how many positive detections in the buffer are needed to count an occupant.

After a testbench using several offline video sequences, the best results we could get showed a video processing speed between 10 to 15 fps and a percentage of right guess from 60 to 80% aproximately. In spite of these interesting results, Viola and Jones face detector plus standard frontal and profile faces databases showed some problems since there are a lot of frames in which the passengers' heads were so much rotated, faces were partially occluded or shadows and illumination changes affected faces appearance. Furthermore, images are radially distorted, so face deformations deteriorate the detection process. In fact, undistortion does not help either, because the interpolation of pixels near the edges of the frames distorts even more the occupants' faces.

There are some works for pose and head rotation estimations such as [17] [18], but we do not focus our work in head pose estimation. In order to compare the detection performance of Viola and Jones algorithm against Bag of visual words, we make a single-frame processing to obtain the *confusion matrix* representation where percentage values concerning true and false, positives and negatives classifications are presented. Table I exhibits the confusion matrix for the Viola and Jones face detector over the 1,400 image test set.

Low face detection values presented in Table I and in Fig. 7 are due to the subset of 700 difficult image

TABLE I

VIOLA AND JONES FACE DETECTOR. CONFUSION MATRIX VALUES

		True category	
		Background	Face
Estimated category	Background	100%	91,08%
	Face	0%	8,86%

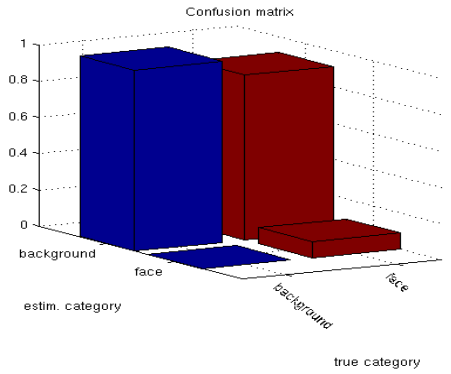


Fig. 7. Viola and Jones face detector. 3D representation of confusion matrix.

patches that contain faces to detect. At the beginning of Section V, we mentioned the main features of the selected images in the dataset, i.e. images that include, among others, illumination changes, face occlusions, shadows and rotated faces. The standard databases of faces that we used for the Viola and Jones face detector are not trained for those difficult images. However, we did not build new databases of faces because of the large amount of images that are needed, e.g. around 1,000-10,000 positive image cases. Fig. 8 contains some samples of images in which Viola and Jones face detector fails.

Fig. 8. False negatives. Difficult images containing faces missclassified by Viola and Jones face detector.

B. Bag of Visual Words

Bag of visual words method is based on image categorization in different classes. In the training stage, this method selects a number of image patches from each category randomly in order to build the vocabulary tree of visual words. After that, the testing stage takes the rest of the image patches, which are classified by searching in the vocabulary tree.

Fig. 9 shows the confusion matrix associated to the classification of the 1,400 image patch dataset. For this test, the method uses 200 training images from each subset.

TABLE II

BAG OF VISUAL WORDS. CONFUSION MATRIX VALUES

		True category	
		Background	Face
Estimated category	Background	81,08%	7,62%
	Face	18,92%	92,38%

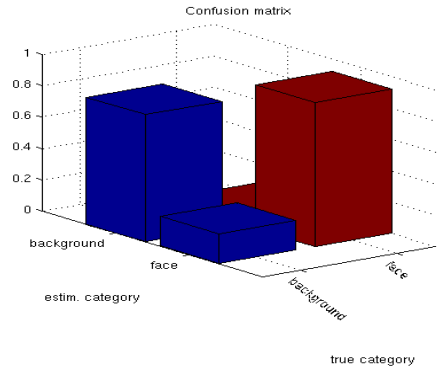


Fig. 9. Bag of Visual Words. 3D representation of confusion matrix.

Table II exhibits much better results than for the Viola and Jones face detector. Near 93% of image patches containing faces are correctly classified as belonging to *face category*. On the other hand, true negatives, i.e. background image patches classified as background, has a ratio of 81%. We have perfomed additional tests using 100 and 150 training images and we made several repetitions due to those images being randomly choosen by the algorithm. We observed very similar results comparing the confusion matrixes on each test, around $\pm 10\%$ deviations at maximum considering the diagonal values shown in Table II. Then, we conclude that the probability of correct classification is high and there are few cases in which this method fails. Some examples of failures are included in Fig. 10. The vast majority of *false positives* are due to partial illuminated areas or partial occluded seats because of the adjacent occupant. On the other hand, *false negatives* are the lowest value in the confusion matrix and some of these missclassified images contain a user wearing sunglasses.

Fig. 10. Bag of Visual Words. Missclassified images patches.

We have not computed the processing time of this method yet. Nevertheless we have taken into consideration the estimations included in [3] that refers to 0.2s in feature extraction and 25ms on vocabulary tree query for a 640x480 resolution video frame.

VI. Conclusions and Future Works

A. Conclusions

This paper has presented preliminary results of two methods that have been applied to an occupant monitoring system that requires and onboard vehicle camera and processing resources in the infrastructure side. The first approach uses Viola and Jones face detector and two standard databases of profile and frontal faces, so the rejection cascade of classifiers is already trained. On the other hand, Bag of visual words method builts a vocabulary tree choosing a random subset of images from our test set. Classification results are much better in the second approach and furthermore, it uses less images for training than in the first one.

Bag of visual words method optimizations depend on the correct modelling of the dataset and the algorithm implementation, as long as Viola and Jones face detector depends on several parameters that have been set empirically.

Finally, we should remark that the low values in Table I are not caused by a misuse of Viola and Jones face detector. The variability in faces appearance because of radial distortion, illumination changes and natural heads movements deteriorate the detection performance of the Viola and Jones algorithm.

B. Future Works

As future works we are interested in the improvement of Bag of visual words method, hence we propose some future guidelines.

- Comparisons of different scale invariant descriptors, such as SURF [19] or M-SURF [20].
- Addition of new classes that can happen in common driving conditions (e.g. babies' seats, pets, etc.).
- Further analysis and optimizations about image dataset for training and testing, which will let better estimations about reliability and performance of this method.
- Implementation of a multi-frame approach in order to process real-time videos, which will reduce considerably the total number of false positives.
- Inclusion of some additional indicators, like "open door" and "fasten seat belt", which will help develop a more robust and reliable occupant monitoring system.
- Further research to check the performance of this method at nighttime using infrarred illumination.
- Integration with the communication module and systems at the infrastructure side in order to perform traffic control tasks in HOV lanes and parking lots.

VII. Acknowledgements

This work was supported in part by the Spanish Ministry of Education and Science (MEC) under grant TRA2008-03600 (DRIVER-ALERT Project) and by the multinational corporation FICOSA INTERNATIONAL.

References

[1] McCormick Rankin Corporation, "Automated Vehicle Occupancy Monitoring Systems for HOV/HOT Facilities," 2004, ontario Ministry of Transportation. Canada.

[2] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[3] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[4] B. F. M. B. N. P. I. Pavlidis, P. Symosek, "Automatic detection of vehicle occupants: the imaging problem and its solution." *Machine Vision and Applications*, vol. 11, no. 6, pp. 313–320, 2000.

[5] L. M. L. J. R. Tyrer, "An optical method for automated roadside detection and counting of vehicle occupants," in *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 222, no. 5, January 2008, pp. 765–774.

[6] P. Consultancy, "High occupancy vehicle monitoring (HOVMON)," http://www.photonicsconsultancy.co.uk/hovmon.htm.

[7] "Vehicle Occupancy Ltd." http://www.vehicleoccupancy.com.

[8] J. Wikander, "Automated vehicle occupancy technologies study: Synthesis report," Texas Transportation Institute. The Texas A&M University System, Tech. Rep., 2007.

[9] Y. yang, G. Zao, and J. Sheng, "Occupant Pose and Location Detect for Intelligent Airbag System Based on Computer Vision," *International Conference on Natural Computation*, vol. 6, pp. 179–182, 2008.

[10] A. Makrushin, M. Langnickel, M. Schott, C. Vielhauer, J. Dittmann, and K. Seifert, "Car-seat occupancy detection using a monocular 360° nir camera and advanced template matching," in *DSP'09: Proceedings of the 16th international conference on Digital Signal Processing*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 1044–1049.

[11] S. Wender and O. Loehlein, "A cascade Detector approach applied to vehicle Occupant Monitoring with an Omnidireccional camera," in *IEEE Intelligent Vehicles Symposium (IV)*, University of Parma, Italy, 2004.

[12] J. Wu, S. C. Brubaker, M. D. Mullin, and J. M. Rehg, "Fast asymmetric learning for cascade face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 3, pp. 369–382, 2008.

[13] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proc. European Conf. Computational Learning Theory*, 1995, pp. 23–37.

[14] B.Fulkerson, A. Vedaldi, and S. Soatto, "Localizing objects with smart dictionaries," in *Eur. Conf. on Computer Vision (ECCV)*, 2008.

[15] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] *Some Methods for classification and Analysis of Multivariate Observations.*, Berkeley, University of California Press, 1967.

[17] M. Jones and P. Viola, "Fast multi-view face detection," Mitsubishi Electric Research Laboratories, Tech. Rep., 2003.

[18] Y. L. Chang Huang, Haizhou AI and S. LAO, "Vector Boosting for Rotation Invariant Multi-View Face Detection," *ICCV*, 2005.

[19] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[20] M. Agrawal, K. Konolige, and M. R. Blas, "CenSurE: Center Surround Extremas for realtime feature detection and matching," in *Eur. Conf. on Computer Vision (ECCV)*, 2008.