Evolving GMMs for road-type classification

Mahmud Abdulla Mohammad, Ioannis Kaloskampis and Yulia Hicks
School of Engineering, Cardiff University
Queen's Buildings, The Parade, Cardiff, UK
Email: {MohammadMA, KaloskampisI, HicksYA}@cardiff.ac.uk

***Abstract*—In this paper, a new online vision-based road-type classification method is proposed. The method uses video captured by a single video camera and takes into account the visual information of the whole scene by segmenting the video frames into temporally consistent frame segments. To this end, we use a video segmentation algorithm based on evolving Gaussian mixture models (GMMs). Our method consists of two stages. In the first stage, we build a priori statistical models of different road types, one model per road type under consideration. For this purpose, we use GMMs produced by the video segmentation algorithm applied to the training video data offline. In the second stage, new video frames are segmented and classified into one of several possible road types on the basis of the Bhattacharyya distance between the Gaussians produced from the new video frame and the Gaussians from the a priori models representing the different road types. Experimental results on real-world data indicate that our method outperforms the state of the art method in this area in both classification accuracy per road type and overall classification accuracy.**

I. Introduction

Vision-based road-type classification can be described as the process of specifying road types based on the video content of the scene. This task is an important step towards road scene understanding, which is required in a variety of applications of situational awareness and fully or semi-automated driving [1]. In such applications, exploiting domain knowledge information is key. However, extracting domain knowledge information from the perception of the road environment is a major issue in autonomous systems [2]; this needs high-quality image/video processing methods [3].

Over the last three decades, many research contributions have been made to the area of visual navigation [4]; nonetheless, building robust methods is still an open problem [2]. In recent years, there has been a considerable amount of research in this area based on different types of sensors, but in terms of the cost and richness of information, using a monocular camera is preferable [5]. Examples of work in this area include road environment classification [3] [1], road detection [6] [7], road marking [8], road sign detection and recognition [9], on-road sign analysis [10], off-road environment classification [11], and highway lane detection [12].

The work presented in [1] [3] focuses on the problem of road-type classification, as in our work. There are three main steps in each of these approaches: region selection, feature extraction and preparation, and classification. Both methods select three subregions of interest from the frames of the road video sequences—road, road edge, and road side—but use different features and classifiers in the second and third steps. The method in [1] extracts colour, texture, and edge-derived features and applies k-nearest neighbor (k-NN) and artificial neural network (ANN) classification approaches, whilst the method in [3] extracts Gabor texture features and uses the random forests classifier [13]. The method in [3] achieved higher accuracy classification than the method in [1], which we attribute to the selection of strong features.

In both methods [1] [3], three subregions were selected as the interest regions for the driving environment: *road, road edge, and road side*. The properties of these three regions are captured and used as the key information during classification. However, there is no guarantee that the subregions capture all key information. In addition, there are specific cases in which the above regions are not likely to contain the key information, such as, for example, when the car turns left or right or is driven on a rough road. To overcome such issues it is necessary to take into account all regions in the frame. One way to achieve this is to use an online video segmentation method and then compare the detected segments to those usually found in certain types of roads.

In this article, we propose a method to classify road types on the basis of the data obtained using a monocular camera. We consider a four-class problem, as in [1], where the classes are motorway, off-road, trunk road, and urban road. Our method consists of two stages. The first stage is building a statistical model for each road type offline. The second stage is the online classification of new video frames. The first stage can be divided into two steps. In the first step, the training frames are segmented using an evolving Gaussian mixture model (GMM) [14]. In the second step, we create a model for each road type from all the Gaussians taken from video sequences illustrating this road type. The classification stage can be divided into two steps. In the first step, an evolving GMM is created for the new frame. We then use the Bhattacharyya distance [15] to find the distance between the Gaussians from the new frame and the models obtained from first stage, which allows us to classify each of the new Gaussians as belonging to one of the road types. In the second step, the road type confidence score is calculated based on the size of the segment corresponding to each classified Gaussian. Experimental results on real-world data indicate that our method outperforms the previous state of the art method in this area in terms of classification accuracy.

The remainder of the paper is organised as follows: In Section II, we present an overview of an evolving GMM. In Section III, we discuss the models of different road types. In Section IV, we give a detailed description of the classification approach. Section V provides experimental results, and Section VI concludes the article.

978-1-4799-7800-7/15/$31.00 ©2015 IEEE

II. Online Video segmentation

We build the road-type models using the evolving GMM algorithm from [14]. In this section we give a high-level overview of this algorithm and justify its use. As we mentioned in the Introduction, we would like our model to store information from all regions within a video frame. At the same time, as we want our method to work in real time, it should handle this information efficiently.

Several videos are used in order to build the model for each road type. Each of these videos is processed as follows: For every frame in the video, we extract visual features from each of its pixels. We then build a GMM using the features of the frame, and after building the GMM, all features extracted from the pixels are discarded. Thus, each frame is represented by a GMM rather than its pixel features, which saves a significant amount of computer storage space and memory (in our case study, we estimate that the GMM representation of a video frame takes up only 0.03% of the memory that its pixel features would require).

The representation of a video sequence could simply be the concatenation of the components of the GMMs, which were built on all frames of the sequence. However, this would lead to a complex model that would include a large number of overlapping components. The evolving GMM algorithm from [14] overcomes this problem as follows: After concatenating the GMM components built on all video frames, it merges the components using a modified version of the expectation-maximisation algorithm. This process results in a compact *merged* model with no overlapping components. The size of this merged model is similar to that of a simple GMM generated on a single frame. see [14] for more details regarding the merging process.

The final model for a road type results from the concatenation of all merged models that were built on video sequences illustrating this road type. To segment a video frame, each pixel in the frame is attributed to a segment according to its probability as estimated with the pdf of the final model. The chosen method is suitable for online applications; moreover, it provides consistent segmentation by preserving long-term information throughout the frames.

III. Building the road-type model

In this section we describe the process we follow in order to build a model for each road type. This process is illustrated in Figure 1. For each road type, i, we select a set S_i of m image sequences illustrating the road type i. The set S_i is given by the equation

$$S_i = \{I_i^{(1)}, I_i^{(2)}, ..., I_i^{(m)}\} \tag{1}$$

where $I_i^{(n)}, n \in \{1, 2, ..., m\}$ is an image sequence of road type i. We then extract visual features from every frame of each image sequence in set S_i. Following [16] [14], we achieve this by representing each pixel in each frame with a five-dimensional vector that includes the pixel's colour descriptor in the Lab colour space and the pixel's spatial coordinates. We denote with $F_i^{(n)}$ the feature representation of an image sequence $I_i^{(n)}$ and thus obtain the set of feature representations $S_i{}'$ as

$$S_i{}' = \{F_i^{(1)}, F_i^{(2)}, ..., F_i^{(m)}\} \tag{2}$$

We then apply the evolving GMM algorithm from [14] to all the feature representations of $S_i{}'$; thus, each element of $S_i{}'$ becomes a GMM. Finally, we concatenate all resulting GMMs into a unified model. The model M_i for road type i is given by the formula

$$M_i = \{L_{ik}\}_{k \in \{1,2,..,N_i\}} \tag{3}$$

where L_{ik} is the k^{th} Gaussian in M_i and N_i is the total number of Gaussians in M_i. In this work, we consider four road types, off road, motorway, urban road, and trunk road, following the suggestion of [1].

In the next section we explain how an input frame is assigned to a road type using our model.

IV. Classification

We assign an input frame f to a road type M_i by estimating its proximity to each road-type model. We first build the model for each frame, M_f, which is a GMM estimated on frame f given by the equation:

$$M_f = \{G_{fj}\}_{j \in \{1,2,..,N_f\}} \tag{4}$$

where G_{fj} is the j^{th} Gaussian and N_f the total number of Gaussians in model M_f.

Our next step is to estimate the distance between each Gaussian from the segmented frame f and the models, using the Bhattacharyya distance [15], which is defined as

$$B\left(G_{fj}, L_{ik}\right) = \frac{1}{8}\left(\mu_{fj} - \mu_{ik}\right)^T \Sigma^{-1} \left(\mu_{fj} - \mu_{ik}\right) + \frac{1}{2}\log\left(\frac{\det\Sigma}{\sqrt{\det\Sigma_{fj}\det\Sigma_{ik}}}\right) \tag{5}$$

where $B\left(G_{fj}, L_{ik}\right)$ is the Bhattacharyya distance between the j^{th} Gaussian of the GMM of f and the k^{th} Gaussian of model M_i. We denote with (μ_{fj}, Σ_{fj}) and (μ_{ik}, Σ_{ik}) the means and covariances of the j^{th} Gaussian in f and the k^{th} Gaussian in model M_i, respectively. For Σ it is the average between Σ_{fj} and Σ_{ik}.

We then find the minimum distance between the j^{th} Gaussian in f and the Gaussians in model M_i. This distance, denoted with β_{fij}, is estimated as

$$\beta_{fij} = min\left\{B\left(G_{fj}, L_{ik}\right)\right\} \tag{6}$$

We classify the Gaussians on the basis of the distances. We consider four road types; thus, the classification has four possible decision outcomes. The decision is given by the equation

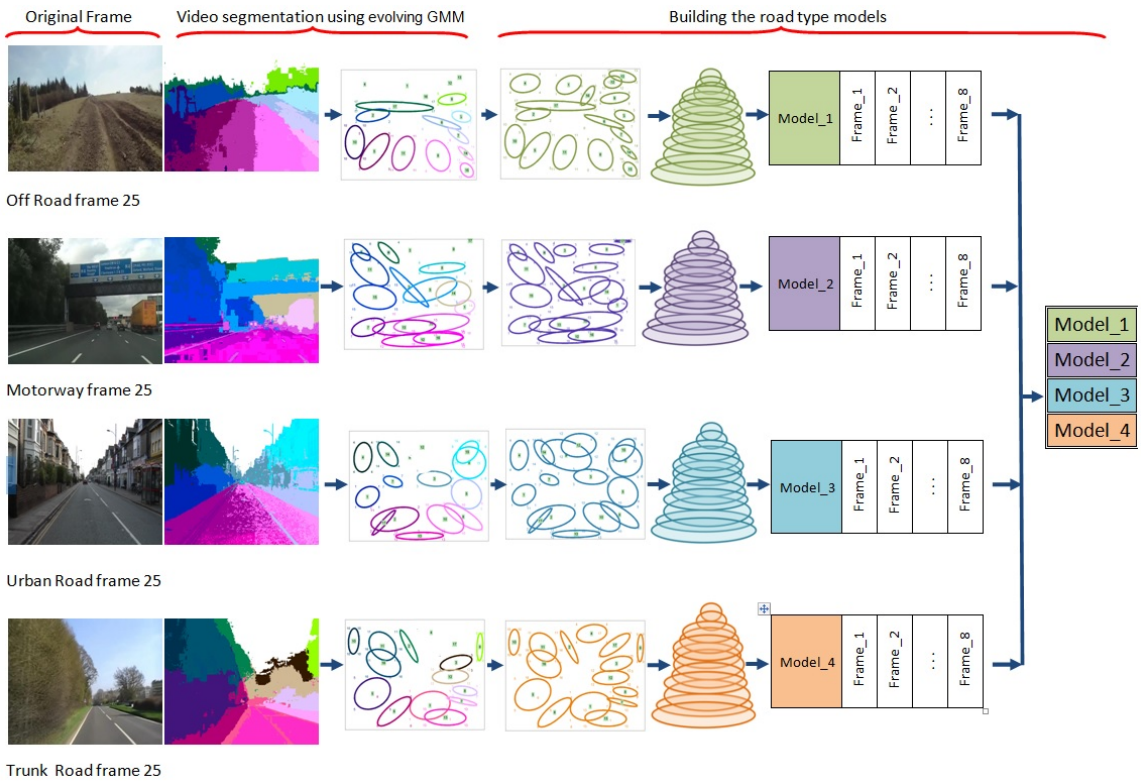

Fig. 1. The process of building the road-type models.

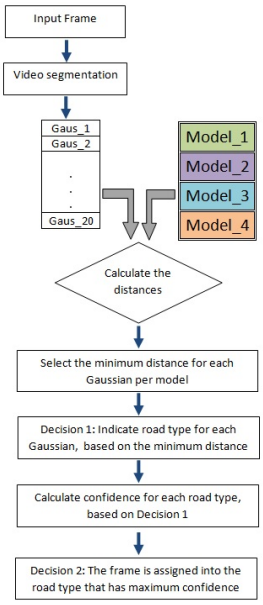

Fig. 2. The pipeline of the classification process.

$$D_{fj} = \arg\min_i \|\beta_{fij}\| \tag{7}$$

where D_{fj} is the classification outcome for the j^{th} Gaussian of f. Equation 7 returns the road type assigned to each Gaussian.

Having classified the Gaussians of f, we estimate the road-type confidence score C_{fi} for road type i, which we define as the percentage of pixels in f that vote for this road type:

$$C_{fi} = \sum_{j=1}^{N_f} size(R_{fj}) \times (D_{fj} = i) \tag{8}$$

where R_{fj} is the segmented region in f that corresponds to the j^{th} Gaussian of G_{fj}. The final decision F_f is made by selecting the road type that maximises the confidence score:

$$F_f = \arg\max_i \|C_{fi}\|. \tag{9}$$

V. Experimental Results

We built the model for each road type using eight videos of 25 frames each. Thus, each model is built using 200 frames. The videos used for the urban road model were taken from [17], while the videos for the rest of the road types were taken from YouTube. The initial resolution of the videos varied, and the frame rate was between 25 fps to 30 fps. We re-sized the resolution of all video frames to 640 * 480. All videos were captured from right-hand drive vehicles and correspond to the drivers perspective, with legal and safety speed limits for each road type. For testing, we used 800 video frames illustrating each road type that were collected in a similar way as the videos mentioned above. These frames were not used when building the road-type models.

We also implemented the state of the art method from [3] to benchmark the performance of our method. We used the same training and testing datasets as above. The method uses random forests [13] for classification; however, we gradually

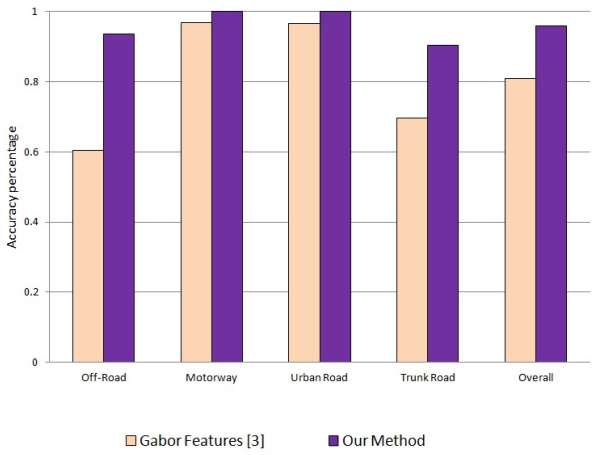

Fig. 3. Road-type classification results using different methods.

TABLE I. CLASSIFICATION RESULTS IN TERMS OF PERCENT CLASSIFICATION ACCURACY

Road types and methods	Our method	State of the art [3]
Off Road	**93.6%**	60.4 $\pm$3.47%
Motorway	**100%**	96.8 $\pm$1.00%
Urban Road	**100%**	96.6 $\pm$0.49%
Trunk road	**90.3%**	69.7$\pm$5.05%
Overall	**96%**	80.9 $\pm$2.5%

increased the number of trees in the forest and finally used 100 trees, which gave the highest classification accuracy. For more than 100 trees, the gain in classification accuracy was insignificant. We ran the random forests classifier 10 times and recorded the mean and standard deviation of the classification accuracy.

The classification results in terms of percent classification accuracy for both methods are presented in Table I and Figure 3. Our method achieves higher classification accuracy than the method from [3] for each road type individually and, consequently, higher overall classification accuracy. The difference between the two methods is more evident in the classification of the off-road environment. Our method achieves 93.6% classification accuracy for this road type, while the accuracy for [3] is 60.4%. This is due to the fact that [3] extracts its features from three predefined subregions in the video frame. However, there is no guarantee that the key information of the scene is always contained within these regions. Since our method collects features from the entire scene, it is expected that in environments where the scenery is more variable, such as in the off-road case, our method will achieve higher classification accuracy.

VI. CONCLUSION

In this work we proposed a method for classifying road types based on video segmentation and the evolving GMMs. All information from the visual content of the scene was used without giving any priority to spatial or perceptual areas of the scene. We considered a four-class problem with four different road types. For testing and comparison with the previous state of the art method in [3], we selected several video sequences of different road types each comprising several hundred frames. We implemented both our method and the method in [3] and split the above dataset into training and testing parts, respectively.

The experimental results demonstrated that our method outperformed the method in [3] in both classification accuracy per road type and overall classification accuracy. We attribute the results to using the information from all areas of the frames. However, the state of the art method in [3] is faster than our method. Our method is online, and there is room for optimisation. Future work will investigate the optimisation of our method as well as testing on more datasets.

REFERENCES

[1] I. Tang and T. Breckon, "Automatic road environment classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 476–484, June 2011.

[2] A. Miranda Neto, A. Correa Victorino, I. Fantoni, and J. Ferreira, "Real-time estimation of drivable image area based on monocular vision," in *IEEE on Intelligent Vehicles Symposium (IV)*, June 2013, pp. 63–68.

[3] L. Mioulet, T. Breckon, A. Mouton, H. Liang, and T. Morie, "Gabor features for real-time road environment classification," in *IEEE International Conf. on Industrial Technology (ICIT)*, Feb. 2013, pp. 1117–1121.

[4] F. Bonin-Font, A. Ortiz, and G. Oliver, "Visual navigation for mobile robots: A survey," *J. Intell. Robotics Syst.*, vol. 53, no. 3, pp. 263–296, Nov. 2008.

[5] W. Liu, L. Zuo, H. Yu, H. Yuan, and H. Zhao, "Obstacle detection based on multiple cues fusion from monocular camera," in *16th International IEEE Conf. on Intelligent Transportation Systems - (ITSC)*, Oct. 2013, pp. 640–645.

[6] J. Alvarez and A. Lopez, "Road detection based on illuminant invariance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 184–193, Mar. 2011.

[7] A. Broggi and S. Berté, "Vision-based road detection in automotive systems: A real-time expectation-driven approach," *J. Artif. Int. Res.*, vol. 3, no. 1, pp. 325–348, 1995.

[8] A. Kheyrollahi and T. P. Breckon, "Automatic real-time road marking recognition using a feature driven approach," *Mach. Vision Appl.*, vol. 23, no. 1, pp. 123–133, Jan. 2012.

[9] G. Piccioli, E. D. Micheli, P. Parodi, and M. Campani, "Robust method for road sign detection and recognition," *Image and Vision Computing*, vol. 14, no. 3, pp. 209 – 223, 1996.

[10] M. Eichner and T. Breckon, "Integrated speed limit detection and recognition from real-time video," in *IEEE on Intelligent Vehicles Symposium*, June 2008, pp. 626–631.

[11] P. Jansen, W. van der Mark, J. van den Heuvel, and F. Groen, "Colour based off-road environment and terrain type classification," in *in Proceedings IEEE on Intelligent Transportation Systems*, Sept. 2005, pp. 216–221.

[12] J. Melo, A. Naftel, A. Bernardino, and J. Santos-Victor, "Detection and classification of highway lanes using vehicle motion trajectories," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 2, pp. 188–200, June 2006.

[13] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] I. Kaloskampis and Y. Hicks, "Estimating adaptive coefficients of evolving gmms for online video segmentation," in *6th IEEE International Symposium on Communications, Control and Signal Processing (ISCCSP)*, May 2014, pp. 513–516.

[15] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, Feb. 1967.

[16] J. Goldberger and H. Greenspan, "Context-based segmentation of image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 463–468, 2006.

[17] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recogn. Lett.*, vol. 30, pp. 88–97, Jan. 2009.