Optical Engineering 49(10), 107201 (October 2010)

# Unsupervised video-based lane detection using location-enhanced topic models

**Hao Sun**
National University of Defense Technology
School of Electrical Science and Engineering
47 Yanwachi Street
Changsha, Hunan 410073, China
E-mail: clhaosun@gmail.com

**Cheng Wang**
**Boliang Wang**
Xiamen University
Department of Computer Science
School of Information Science and Technology
Xiamen, Fujian 361005, China

**Naser El-Sheimy**
University of Calgary
Department of Geomatics Engineering
2500 University Drive Northwest
Calgary, Alberta T2N 1N4, Canada

**Abstract.** An unsupervised learning algorithm based on topic models is presented for lane detection in video sequences observed by uncalibrated moving cameras. Our contributions are twofold. First, we introduce the maximally stable extremal region (MSER) detector for lane-marking feature extraction and derive a novel shape descriptor in an affine invariant manner to describe region shapes and a modified scale-invariant feature transform descriptor to capture feature appearance characteristics. MSER features are more stable compared to edge points or line pairs and hence provide robustness to lane-marking variations in scale, lighting, viewpoint, and shadows. Second, we proposed a novel location-enhanced probabilistic latent semantic analysis (pLSA) topic model for simultaneous lane recognition and localization. The proposed model overcomes the limitation of a pLSA model for effective topic localization. Experimental results on traffic sequences in various scenarios demonstrate the effectiveness and robustness of the proposed method.  [DOI: 10.1117/1.3490422]

Subject terms: lane detection; shape descriptor; unsupervised learning; topic models.

Paper 100226RR received Mar. 21, 2010; revised manuscript received Jul. 12, 2010; accepted for publication Aug. 5, 2010; published online Oct. 28, 2010.

## 1 Introduction

In modern driver-assistance systems, the environment perception plays a decisive role in order to evaluate the current traffic scene. The early and reliable detection of traffic lanes and road users determines the ability of integrated systems to warn the driver in dangerous situations and contributes to road safety. Among others, video sensors are the subject of current research in the lane-detection context, because video sensors enable sophisticated image-processing algorithms and keep the costs to an affordable limit. This paper aims at developing an automatic algorithm for lane detection in video sequences observed by uncalibrated moving cameras.

Detecting lanes is a challenging task owing to their variability, ambiguity, and the wide range of illumination and scale conditions that may apply. A robust algorithm needs to detect the circular reflector, solid lines, and segmented line markings under varying lighting and road conditions, and be able to deal with challenging scenarios, such as lane curvature, worn lane marking, lane changes, and emerging, ending, merging, and splitting lanes.

There has been active research on lane detection in the literature, and a wide variety of algorithms of various representation, including edge points grouping, fixed-width line pairs, and deformable template model, have been proposed. Generally, previous research on lane detection in video sequences can be classified into four categories: edge-based methods, model-based methods, perspective transformation-based methods, and Hough transform-based methods. Edge-based methods detect lane markings by edge-point extraction. However, it is not possible to select a threshold that eliminates noise edges without eliminating many of the interesting lane edge points in many road scenes. Model-based methods often maintain an appropriate geometrical road model and extracts lane features on the road. But it is uneasy to use geometrical models and difficult to detect complex road features. In addition, the computation is complex. The difficulty of perspective transformation methods is the calibration of cameras. The Hough transform algorithm has good ability to overcome the noise, but the selection of the parameters is difficult and the computations are very complicated. In particular, Kreucher and Lakshmanan[1] presented a lane-finding in another domain (LANA) system to extract lane markings in a frequency domain and to detect lanes with a deformable template. The lane detection in Ref. 2 uses a sobel operator for edge detection, and to handle fine and coarse structures, an image pyramid is constructed. Danescu et al.[3] presented an approach based on stereo cameras and distance-dependent 1-D edge filters. Lane detection is done by means of a dark-light-dark (DLD) transition, where a gradient pair must have similar magnitude but an opposite sign. For more recent development of the field, the interested reader can refer to the work of McCall and Trivedi.[4]

Most of the previous methods are based on various additional assumptions and need some prior information or hypotheses.[5] In this paper, we propose a generative graphical model approach to recognize and localize lane markings in videos, taking advantage of the robust representation of the bag-of-visual-words component and an unsupervised learning algorithm. In the context of our problem, unsupervised learning is achieved by obtaining lane model parameters from unsegmented and unlabeled video sequences based on topic models. We advocate the use of an unsupervised learning setting because it opens the possibility to use the increasing amount of available video data without the expense of detailed human annotation.

0091-3286/2010/$25.00 

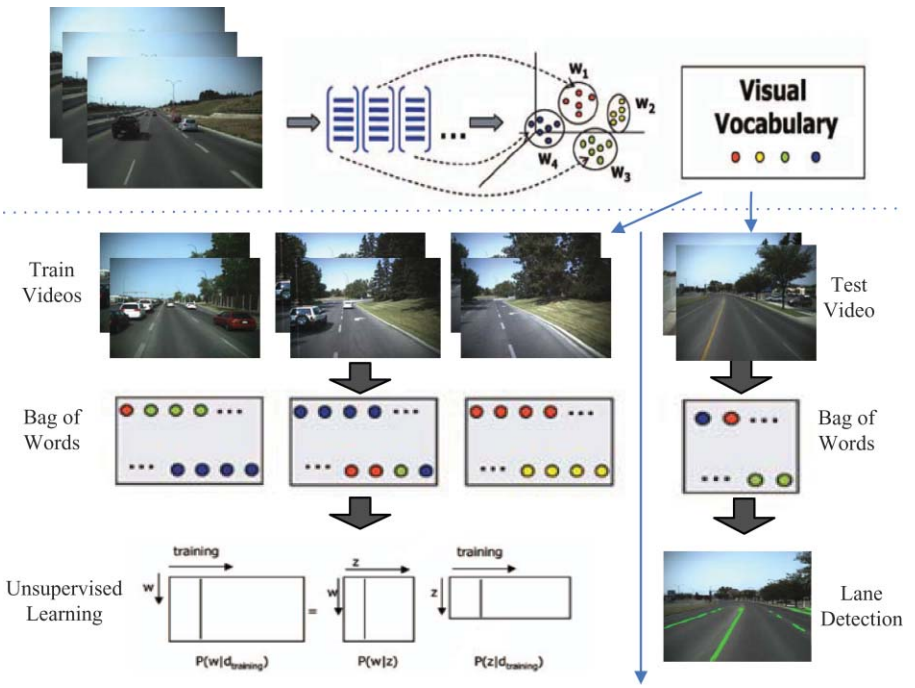



**Fig. 1** Flow diagram of the proposed algorithm.

Given a collection of unlabeled video sequences, our goal is to automatically learn the lane models and to apply the learned model to perform lane recognition and localization in the new video sequences. Figure 1 shows the flow diagram of the proposed method. To represent video sequences, we first extract local patches using the maximally stable extremal region (MSER) detector.[6] These local patches, described by both appearance and shape descriptor, are then clustered into a set of visual words. Probability distributions and intermediate topics are learned automatically using probabilistic latent semantic analysis (pLSA).[7] The learned models can then be used to recognize and localize lane markings in a novel video sequence.

The outline of the paper is as follows. In Section 2, we introduce the bag of visual words model for video representation. Section 3 reviews the pLSA topic model and its limitations in topic localization, and introduces a novel location-enhanced topic model for simultaneous lane recognition and localization. In Section 4 experimental results on traffic sequences in various scenarios are reported. Finally, we conclude the paper in Section 5.

## 2 Video Representation from Visual Words

In this section, we introduce the bag-of-visual-words model used for video representation. First, features are extracted by MSER detector. A modified scale-invariant feature-transform (SIFT)[8] descriptor and a novel shape descriptor defined in local affine frames are then proposed to describe the detected features. Finally, the visual vocabulary is built by quantizing the descriptors and a video is represented by bag of visual words.

The bag-of-words model is a simple assumption used in natural language processing and information retrieval and has been widely used in the computer vision field.[9,10] In this work, we adopt it for video-content representation. In general, there are three main steps for the model: (*i*) Extract local features and obtain their descriptors; (*ii*) quantize the descriptors into a visual vocabulary; and (*iii*) describe the image or video as a collection of visual words.

First, we find a number of local patches to generate the visual words. Candidate patches are determined by running the MSER detector. MSER features have been shown to outperform other affine region detectors on a wide range of test images in a recent evaluation.[11] An extremal region is a connected component of pixels that are all brighter or darker than all the pixels on the region's boundary. The MSER detector finds extremal regions that are stable with respect to the change of intensity thresholds and performs well on images containing homogeneous regions with distinctive boundaries. We rely on the stability of the MSER features to provide repeatable closed contours and robustness to lane-marking variation in scale, lighting, and viewpoint. Figure 2 shows the result of the MSER detector on a frame from an on-road traffic sequence compared to the sobel edge detector and Hough line detector. It can be seen that MSER features are more robust compared to edge points or line pairs for lane-marking extraction.

Two different kinds of complementary descriptors, shape descriptor, and appearance descriptor are then adopted to describe the detected MSER patches (Fig. 3). In our case, color information is discarded. Patches and descriptors are extracted from gray-scale images for process efficiency.

### 2.1 *Local Affine Frame on MSER*

To describe the shape information in an affine invariant way, local affine frames (LAFs) are defined on MSER patches. Several methods for determining LAFs on a MSER can be found in Ref. 12. We use the center of gravity (two constraints) of a region, the symmetric $2\times2$ covariance matrix (three constraints), and dominant gradient orientation of the contour (one constraint) to define a LAF. Each region is represented by a closed polygon constructed from its outer

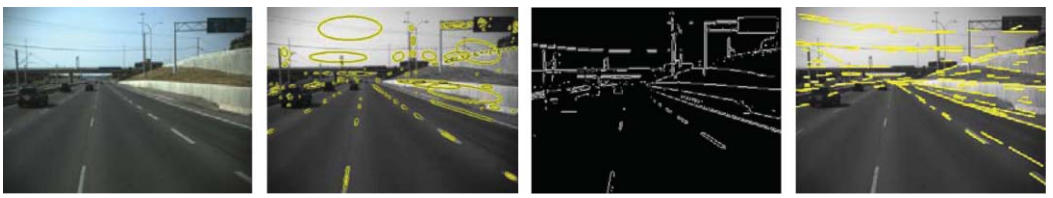



**Fig. 2** Visual feature extraction: (a) A frame from an on-road traffic video sequence, (b) detected MSERs are shown by the yellow ellipses(subset), (c) Sobel edge detection results, and (d) Hough line extraction results.

boundary. Once the covariance matrix is computed, the region shape is normalized so that the covariance matrix of the resulting shape is equal to the identity matrix. Shape normalization, together with the position of the center of gravity of the region, fixes the affine transformation up to an unknown rotation. In order to find reference orientations for the normalized region, gradients on the outer boundary of the region are computed and a histogram of gradient directions is formed. The gradient votes in the histogram are weighted with the gradient magnitude.

### 2.2 *Shape Descriptor in Local Affine Frame*

Any representation of the normalized patches in LAFs is theoretically invariant to affine geometric transformations. Because the region is often homogeneous with distinctive boundaries, we propose a shape descriptor for describing the closed contour. The shape descriptor is a histogram of relative position between the points on region boundary and center of gravity of the region. The procedure to obtain the shape descriptor is as follows: (*i*) Extract the shape contours of the normalized region, (*ii*) construct a coordinate with the region center as its origin, the reference orientation as the axis direction, (*iii*) for each contour point, compute its relative distance $\rho$ and angle $\theta$ to the region center, and (*iv*) count the number of points falling into same bin of direction weighted with the distance. Figure 3 shows the process of constructing shape descriptor.

### 2.3 *Appearance Descriptor Based on Gradient Mirroring*

Mikolajczyk and Schimd[13] compare the performance of descriptors computed for local interest regions, and the results show the SIFT-based descriptors are the best. The SIFT descriptor is computed by partitioning the image region surrounding each detected keypoint into a 4×4 grid of subregions and computing an orientation histogram of eight bins in each subregion. Within each subregion, the gradient orientation of each pixel is entered into the orientation histogram, with weighted vote proportional to the gradient magnitude. A normalized 128-component vector is formed by concatenating the 16 region containers. Although the SIFT descriptor is invariant to linear changes intensity, lane markings in different scenarios often involve nonlinear changes. To improve the robustness for lane-marking detection, gradient mirroring (GM)[14] for appearance descriptor construction is adopted. GM associates antiparallel gradient directions and therefore considers gradient directions in the interval $[0, \pi)$ instead of $[0, 2\pi)$. After GM, eight bins in the interval $[0, \pi)$ are used in each subregion orientation histogram computation. The GM SIFT descriptor has the same length as the original SIFT descriptor and is invariant to contrast reversals. Patch content of the detected features is described using the GM SIFT descriptor.

Normalization of the descriptors makes them robust to contrast changes or scale changes. Local L1 and L2 norms give comparable results. In our experiments, we use the L2 norm to perform a global normalization of the descriptor.

In order to learn the vocabulary of visual words, we consider the set of shape and appearance descriptors corresponding to all detected features in the training data. We used two visual vocabularies, one for the shape descriptor defined in LAF and one for the GM SIFT descriptor. The vocabularies are constructed by clustering descriptors using the *k*-means algorithm and Euclidean distance as the clustering metric. The number of clusters is chosen empirically to maximize detection performance on a manually labeled ground-truth video sequence. The center of each clustering is defined to be a visual word. Thus, each detected feature can be assigned a unique cluster membership such that a video can be

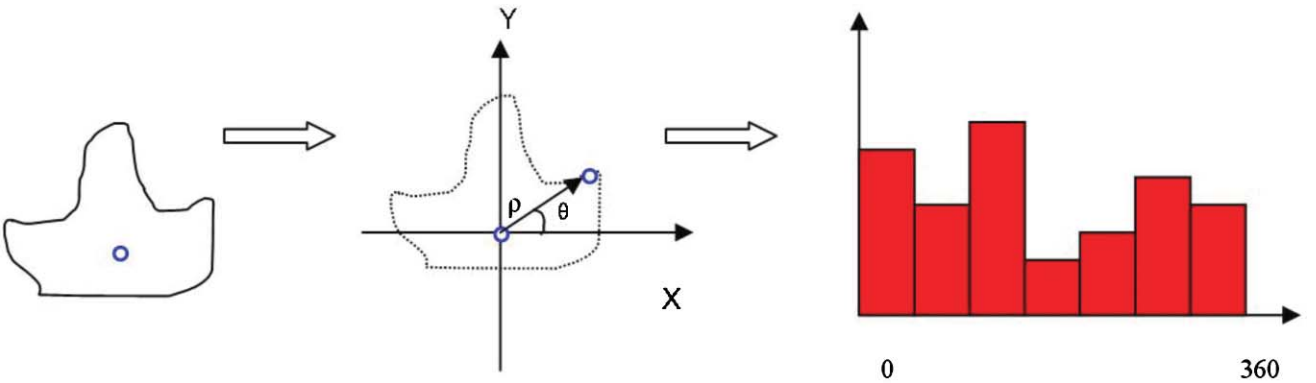



**Fig. 3** Shape descriptor for MSER defined in LAF.

represented as a collection of visual words from the visual vocabulary.

## 3 Lane Detection by Topic Discovery

In this section, we first review the pLSA model and its limitations in topic localization. A novel location-enhanced topic model is then introduced for simultaneous lane recognition and localization.

Topic models, such as pLSA[7] and latent Dirichlet allocation (LDA),[15] have recently been introduced as an alternative over the simple Naive Bayes model for bag-of-words representations. Topic models consider the bag of words as a mixture of several topics. Each document has its own distribution over topics, and each topic is represented as a distribution over words. In text analysis, pLSA is used to discover topics in a document using the bag-of-words document representation. In our case, we would prefer to analyze video sequences instead of text documents; video sequences are summarized as a set of visual words instead of text words, and we seek to discover lane markings instead of text topics.

Following the notations used in the text-understanding community, we have $N$ documents containing words from a vocabulary of size $M$. The corpus of text documents is summarized in a $M$ by $N$ co-occurrence table, where $n(w_i, d_j)$ stores the number of occurrences of a word $w_i$ in document $d_j$. In addition, there is a latent topic variable $z_k$ associated with each occurrence of a word $w_i$ in a document $d_j$.

### 3.1 *pLSA*

The joint probability $P(w_i, d_j, z_k)$ is assumed to have the form of the graphical model shown in Fig. 4(a). Marginalizing over topics $z_k$ determines the conditional probability $P(w_i \mid d_j)$,

$$P(w_i \mid d_j) = \sum_{k=1}^{K} P(w_i \mid z_k) P(z_k \mid d_j), \tag{1}$$

where $P(z_k \mid d_j)$ is the probability of topic $z_k$ occurring in document $d_j$, and $P(w_i \mid z_k)$ is the probability of word $w_i$ occurring in a particular topic $z_k$. pLSA assumes the joint distribution of $d$, $w$, and $z$ can be written as

$$P(d, w, z) = P(d) P(z \mid d) P(w \mid z). \tag{2}$$

### 3.2 *Location-Enhanced pLSA*

pLSA is known for its capability of handling polysemy and has been successfully applied to scene categorization and object recognition.[16] However, one limitation of the pLSA model is its weakness in localizing objects.[17] In order to overcome this limitation, we propose a location-enhanced pLSA (LE-pLSA) topic model for lane detection. The graphical model of LE-pLSA is shown in Fig. 4(b). Let $x$ denote the location of a patch, for a patch in image $d$ with appearance $w$ and location $x$, the joint distribution $P(d, w, x, z)$ has the form,

$$P(d, w, x, z) = P(d) P(z \mid d) P(w \mid z) P(x \mid z), \tag{3}$$

where $P(x \mid z)$ is a spatial distribution that models where a patch with topic $z$ is more likely to occur (e.g., the traffic lane is more likely to be detected on the road rather than other locations). By including location information in the model, the spatial ordering of visual words contributes to the discovery of topic and helps to localizing words of certain topic. The spatial distribution $P(x \mid z)$ uses the same parameters across all documents, and hence, it is a global location model. In our experiments, the spatial distribution is given as a prior.

Fitting the model involves determining the topic vectors $P(w \mid z)$, which are common to all documents, and the mixture coefficients $P(z \mid d)$, which are specific to each document. The goal is to determine the model that gives high probability to the words that appear in the corpus, and a maximum likelihood estimation of the parameters is obtained by maximizing the objective function

$$L = \prod_{i=1}^{M} \prod_{j=1}^{N} P(w_i \mid d_j)^{n(w_i, d_j)}, \tag{4}$$

where $P(w_i \mid d_j)$ is given by Eq. (1). This is equivalent to minimizing the Kullback–Leibler divergence between the measured empirical distribution $\tilde{P}(w \mid d)$ and the fitted model. The model is fitted using the expectation maximization (EM) algorithm as described in Ref. 7.

The posterior is modeled as

$$P(z_k \mid x_l, w_i, d_j) = \frac{P(w_i \mid z_k) P(x_l \mid z_k) P(z_k \mid d_j)}{\sum_{k=1}^{K} P(w_i \mid z_k) P(x_l \mid z_k) P(z_k \mid d_j)}. \tag{5}$$

Once each patch has been assigned to a visual word, we can label the corresponding word $w_i$ with a particular topic by finding the maximum of the posteriors $P(z_k \mid x_l, w_i, d_j)$ over $k$. Thus, we label the regions that support the detected patch, effectively producing a topic localization, which corresponds to the localization of potentially lane markings.

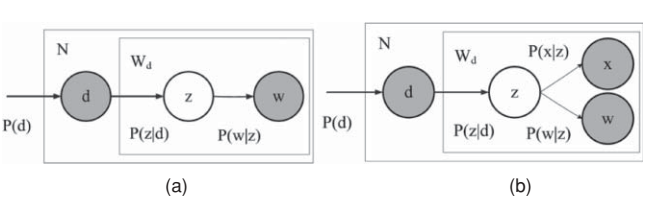
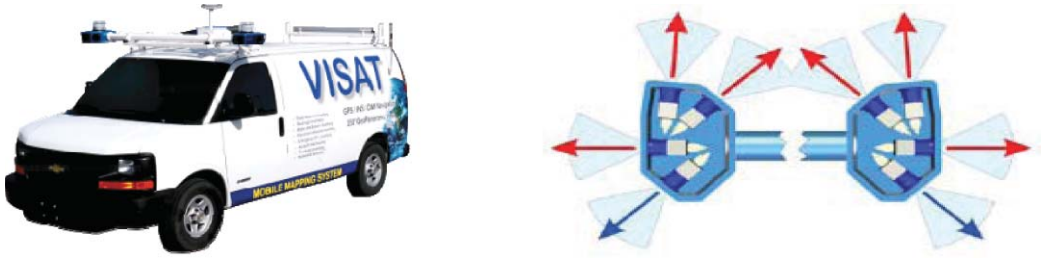



**Fig. 4** Graphical models. Nodes inside a given box indicate that they are replicated the number of times indicated in the top left corner. Filled circles indicate observed random variables while unfilled are unobserved: (a) pLSA graphical model and (b) our location enhanced pLSA graphical model.

## 4 Experimental Results

In this section, experimental results on real-world traffic sequences in various scenarios are reported. Experimental data have been acquired with the VISAT™ mobile mapping system (MMS).[18] VISAT [Fig. 5(a)] was developed at the University of Calgary, and the system's hardware components include a strap-down inertial navigation system, a dual-frequency global positioning system receiver, and multiple digital cameras [Fig. 5(b)]. In our experiments, we use video sequences acquired from the forward-looking cameras.

The data are collected using the VISAT Station Browser developed by Absolute Mapping Solutions Inc., Calgary, Canada. Figure 6 shows the survey route. The original resolution of frames from the VISAT georeferenced video sequence is 1600×1200 pixels. In the experiments, all frames are downsized to 320×240 pixels for process efficiency.




(a) (b)

**Fig. 5** (a) The VISAT$^{TM}$ MMS van and (b) its vision system.


**Fig. 6** Survey route in Calgary used in the evaluation.

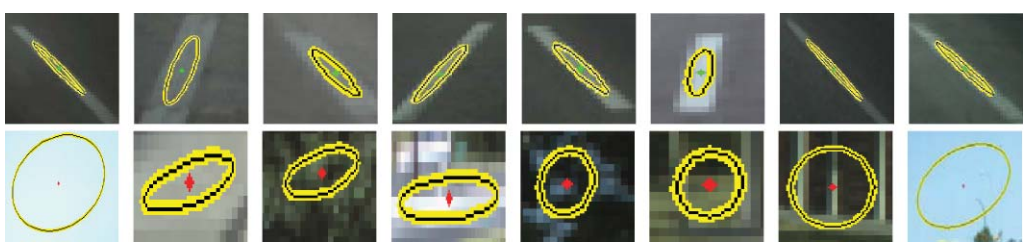
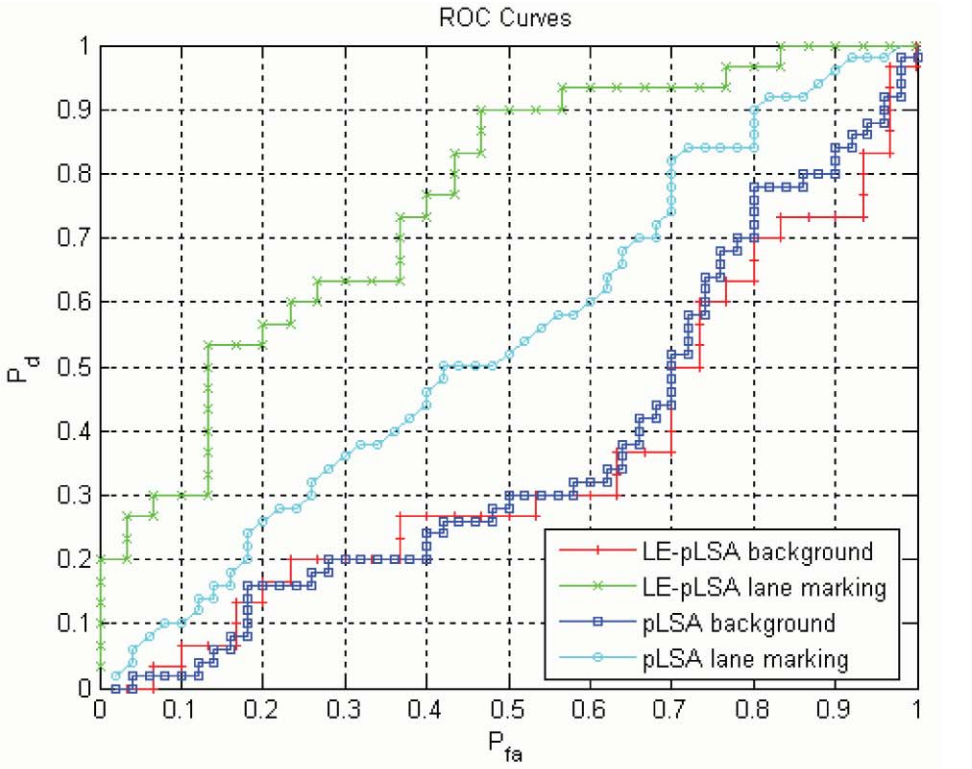
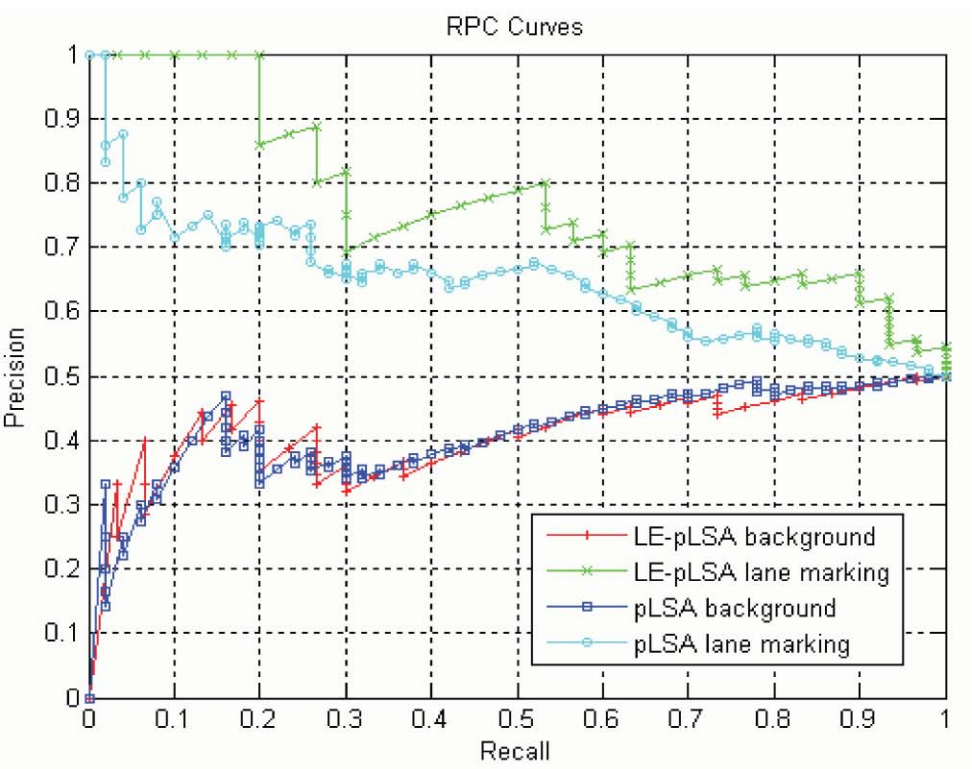


**Fig. 7** Most likely visual words (shown by eight examples in a row) for the two learned topics in our experiments. The first row shows words belonging to lane category; the second row shows words belonging to background category.

**Fig. 8** Selected images from the lane category (first row) and images from the background category (second row) in our data set.




(a)




(b)

**Fig. 9** ROC and RPC curves for images classification using both pLSA and LE-pLSA model fitted on the training data: (a) ROC curves and (b) RPC curves.

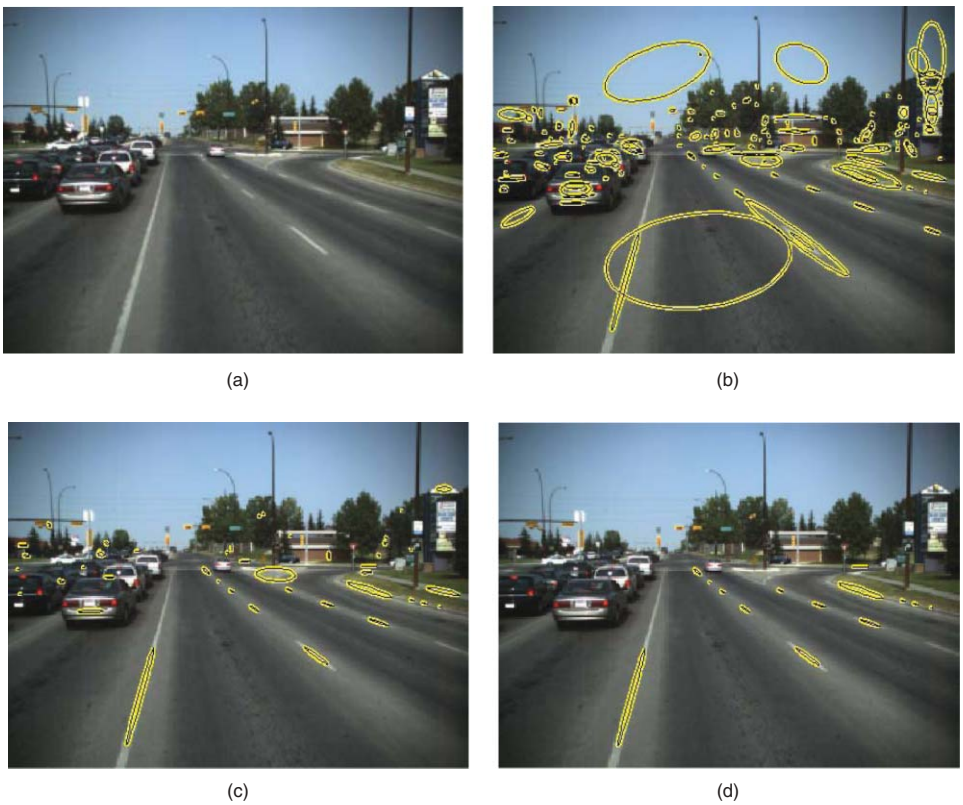



**Fig. 10** Lane detection results using LE-pLSA model: (a) A frame from a road intersection sequence, (b) frame with the detected MSER patches imposed, (c) patches having high probability belonging to the lane marking category, and (d) final localization results.

In the context of unsupervised lane detection, traffic lanes and background clutter are the two topics. We have collected ten unlabeled video sequences in different traffic scenarios for training our model. All descriptors extracted from the training frames are quantized into two visual vocabularies of 300 words. The EM algorithm is initialized randomly and converges in $<100$ iterations. One iteration takes $\sim 1$ s on 1600 frames from the training data with two fitted topics and $\sim 200$ nonzero word counts per frame using our Matlab implementation on a standard 2.66 GHz PC. Visual words that are most probable for the two discovered topics are shown in Fig. 7. Topic discovery analysis cleanly separates the patches into different object classes. Increasing the number of topics in background clutter will help to classify more types of objects (e.g., trees, moving vehicles, buildings, etc.). The most likely words for a topic appear to be semantically meaningful regions.

We use both pLSA and LE-pLSA models to fit the two topics on the training data. To evaluate the performance of the two models, the learned topics are used for classifying new images. Given a new image, the unseen image is projected on the simplex spanned by the learned topic word distributions. A categorization decision is made by selecting the image category that best explains the observation; that is,

$$\text{image category} = \arg\max_k P(z_k \mid d_{\text{test}}). \tag{6}$$

We have collected 2200 images for the evaluation, 1200 images with lane markings (the lane category), and 1000 images without lane markings (the background category). Some selected examples from each category are shown in Fig. 8. In our experiments, the data set has been partitioned into two randomly selected halves; one-half forms the training set, and the other half forms the testing set. The training set consists of 600 randomly selected images from the lane category and 500 randomly selected images from the background category. The rest of the images in the data set forms the testing set for separate performance evaluation.

The receiver operating characteristic (ROC) curves and recall precision curve (RPC) curves for the classification experiments using both pLSA and LE-pLSA are shown in Figs. 9(a) and 9(b), respectively. It can be seen that the proposed LE-pLSA model outperforms the original pLSA model for classification of images from the lane category. For images containing only background clutter, the performances of the two models are similar because spatial locations of patches belonging to the background topic cannot be modeled by a prior distribution.

Figure 10 shows an example of lane-marking localization results using the learned LE-pLSA model. Figures 10(a) and 10(b) show the original frame from a road-intersection sequence and the detected MSER patches. Figure 10(c) shows the patches that have a high probability belonging to a lane-marking category, and Fig. 10(d) is the localization results after spatial filtering. It can be seen that most of the lane markings are correctly detected and accurately localized in the image.

We have tested the proposed algorithm on both on-road video sequences and road-intersection sequences. The road-intersection sequences are more challenging than the on-road sequences because there are more traffic and different kinds of lane markings at the road intersection. Some of the

Frame 21 Frame 39 Frame 46

Frame 65 Frame 93 Frame 122

**Fig. 11** Lane detection results from the on-road sequence.

detection results from the on-road sequence are shown in Fig. 11. Results from the road intersection sequence are shown in Fig. 12.

Figure 11 shows lane-detection results on six frames from the on-road sequence. Note that most near-field lane markings are correctly detected. Far-field lane markings are missed at frames 21, 39, and 46, as few features can be extracted from these markings. Lane markings are correctly detected at frames 65 and 93, despite shadows and lighting variations. The proposed algorithm is robust to complex shadowing, lighting changes from overpasses and tunnels, and road-surface variations. Several false positive are detected at frames 93 and 122.

Figure 12 shows detection results on three frames from the road-intersection sequence. Several false positive are detected at frames 32 and 44. The yellow lane marking at frame 44 is missed. It can be shown that our algorithm can detect most salient lane marking at the road intersection. Through the use of MSER features, lane markings in various shapes can be correctly detected.

To measure the performance of lane-marking detection in video sequences, the following parameters are used:

$$\text{Recall} = N_c/(N_c + N_m), \quad \text{Precision} = N_c/(N_c + N_f), \qquad (7)$$

where $N_c$, $N_m$, and $N_f$ represent the numbers of correctly detected, missed and false-positive lane markings, respectively. The statistics of lane-marking detection results are shown in Table 1, which shows that our algorithm performs better for the case of on-road sequence.

The proposed model can be trained offline, and thus, the learning stage is not time critical. For online lane recognition and localization, the algorithm processing time is closely related to different scenarios as the number of features detected varies under different scenes. Our MATLAB implementation of the algorithm can process ~6 fps on a 2.66-GHz CPU for the on-road sequences. For the road-intersection sequences, the algorithm works a little slower because more features are often detected in the sequences.

## 5 Conclusions

In this paper, a novel unsupervised learning algorithm based on topic models is proposed for video-based lane detection. The major contributions of this paper are as follows: (*i*) The MSER detector is introduced for lane feature extraction. A novel shape descriptor is derived to describe region shapes and a GM SIFT descriptor is presented to capture appearance characteristics. (*ii*) A novel LE-pLSA topic model is proposed for simultaneous lane recognition and

Frame 18 Fame 32 Frame 44

**Fig. 12** Lane detection results from the road-intersection sequence.

**Table 1** Summary results for lane markings detection.

| Sequence Category | Total Frames | $N_c/N_m/N_f$ | Recall | Precision |
|---|---|---|---|---|
| On-road sequence | 1380 | 1932/107/154 | 0.95 | 0.93 |
| Road intersection sequence | 620 | 885/98/139 | 0.90 | 0.86 |

localization. The proposed model overcomes the limitation of the pLSA model for effective topic localization. Experimental results on real-world traffic sequences in various scenarios demonstrate the effectiveness and robustness of our method.

In future research, we will attempt to improve the proposed algorithm for video sequences with severe image clutter and occlusions. Simultaneous detection and recognition of other road users in videos is another future research topic.

## Acknowledgments

The authors thank associate editor Prof. Michael Bove and the anonymous reviewers for valuable comments, which helped improve the clarity of the presentation of this paper. The work was supported by the National Natural Science Foundation of China (Project No. 40971245).

## References

1. C. Kreucher and S. Lakshmanan, "LANA: a lane extraction algorithm that uses frequency domain features," *Robot. Autom.* **15**, 343–350 (1999).
2. F. Samadzadegan, A. Sarafraz, and M. Tabibi, "Automatic lane detection in image sequences for vision-based navigation purposes," in *Proc. of Int. Soc. for Photogrammetry and Remote Sensing Commission V Symp.*, Dresden, pp. 251–257 (2006).
3. R. Danescu, S. Nedevschi, and M. Meineche, "Lane geometry estimation in urban environments using a stereovision system," *Proc. IEEE Transportation Systems Conf.*, Seattle, pp. 271–276 (2007).
4. J. C. McCall and M. M. Trivedi, "Video-based lane estimation and tracking for driver assistance: survey, system, and evaluation," *IEEE Trans. Intell. Transp. Syst.* **7**, 20–37 (2006).
5. Z. Kim, "Robust lane detection and tracking in challenging scenarios," *IEEE Trans. Intell. Transp. Syst.* **9**, 16–26 (2008).
6. J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. of 13th British Machine Vision Conf.*, Cardiff, Wales, pp. 384–393 (2002).
7. T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learning* **42**, 177–196 (2001).
8. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.* **60**, 91–110 (2004).
9. J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.* **79**, 299–318 (2008).
10. X. Wang, X. Ma, and W. E. Grimson, "Unsupervised activity perception in crowed and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 539–555 (2009).
11. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, and J. Matas, "A comparison of affine region detectors," *Int. J. Comput. Vis.* **65**, 43–72 (2005).
12. S. Obdrzálek, "Object recognition using local affine frames," PhD thesis, Czech Technical University (2007).
13. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1615–1630 (2005).
14. A. Kelman, M. Sofka, and C. V. Stewart, "Keypoint descriptors for matching across multiple image modalities and non-linear intensity variations," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, pp. 1–7 (2007).
15. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learning Res.* **3**, 993–1022 (2003).
16. J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. "Discovering objects and their location in images," in *Proc. of 10th IEEE Int. Conf. on Computer Vision*, Beijing, pp. 370–377 (2005).
17. D. Liu and T. Chen, "Unsupervised image categorization and object localization using topic models and correspondences between images," in *Proc. of 11th Int. Conf. on Computer Vision*, Rio de Janeiro, pp. 1–7 (2007).
18. N. El-Sheimy and K. Schwarz, "Navigating urban areas by VISAT— a mobile mapping system integrating GPS/INS/Digital cameras for GIS application," *Navigation* **45**, 275–286 (1999).

**Hao Sun** received his BSc and MSc in National University of Defense Technology, Changsha, China, in 2006 and 2008, respectively. He is currently pursuing a PhD in the School of Electrical Science and Engineering, National University of Defense Technology (NUDT). His research interests include image analysis and understanding, pattern recognition, and information fusion.

**Cheng Wang** received the BSc and PhD in communication and signal processing from the NUDT, Changsha, China, in 1997 and 2002, respectively. He is currently an associate professor in the School of Electronic Science and Engineering, NUDT. He is a professor in the Department of Computer Science at Xiamen University, Fujian, China. He is also the co-chair of the Working Group I/3 "Multi-Platform Multi-Sensor Inter-Calibration" in the International Society for photogrammetry and Remote Sensing (ISPRS). His research interests include image analysis, information fusion, and mobile mapping data processing.

**Boliang Wang** is a professor in the Department of Computer Science at Xiamen University, Fujian, China. He is also a professor of School of Electronic Science and Engineering, NUDT. His area of expertise includes image processing, multisensor integration, and pattern recognition. He has authored over 80 published scientific papers and patents.

**Naser El-Sheimy** is the head of the Department of Geomatics Engineering and leader of the Mobile Multi-sensor Research Group at the University of Calgary, Alberta, Canada. His area of expertise is in the integration of GPS/INS/imaging sensors for mapping and GIS applications with special emphasis on the use of multisensors in mobile mapping systems. He is the chair of the ISPRS Working Group on "Integrated Mobile Mapping Systems," the chair of the special study group for mobile multisensor systems of the International Association of Geodesy, and the chairman of the International Federation of Surveyors working group C5.3 on Integrated Positioning, Navigation, and Mapping Systems.