

# Joint Random Fields for Moving Vehicle Detection

Yang Wang and Getian Ye  
National ICT Australia

## Abstract

This paper proposes a joint random field (JRF) model for moving vehicle detection in video sequences. The JRF model extends the conditional random field (CRF) by introducing auxiliary latent variables to characterize the structure and evolution of visual scene. Hence detection labels (e.g. vehicle/roadway) and hidden variables (e.g. pixel intensity under shadow) are jointly estimated to enhance vehicle segmentation in video sequences. Data-dependent contextual constraints among both detection labels and latent variables are integrated during the detection process. The proposed method handles both moving cast shadows/lights and background illumination variations. Computationally efficient algorithm has been developed for real-time vehicle detection in video streams. Experimental results show that the approach effectively deals with various illumination conditions and robustly detects moving vehicles even in grayscale video.

## 1 Introduction

Moving object detection in video sequences is fundamental in application areas such as visual surveillance, traffic monitoring, human-computer interaction, and video compression. Especially, vehicle detection with stationary camera is an important problem for video based traffic monitoring, which is essential for the measurement of traffic parameters such as vehicle count, speed, and flow. However, accurate detection could be difficult due to the potential variability including shadows or lights cast by moving objects, dynamic background processes, and camouflage [11,15]. Comprehensive modeling of spatiotemporal information within the video scene is a key issue to robustly segment moving objects. Spatial color distribution can be used to characterize background and foreground objects within dynamic scenes [18]. Gradient (or edge) features help improve the reliability of moving object detection [21]. On the other hand, temporal changes of the background can be described by linear processes or statistical distributions according to recent observations [4,23]. In [3,10,19], the recent history of pixel intensity is characterized by a mixture of Gaussians, and the mixture model is adaptively updated for each site to deal with dynamics in background processes. In [2,12], kernel density estimation is employed for adaptive and robust object detection.

Moreover, contextual constraint is an essential element to effectively fuse spatial and temporal information throughout the detection process. Markov random field (MRF) and hidden Markov model (HMM) have been extensively employed to formulate contextual constraints. In [7], HMM is used to impose the temporal continuity constraint on foreground and shadow detection for traffic surveillance. A dynamical framework of topology free HMM capable of dealing with sudden or gradient illumination changes is proposed as well [20]. In addition, spatial smooth constraint is

modeled by MRF in [14,17]. Spatiotemporal MRF involving successive video frames has also been proposed for robust detection and segmentation of moving objects [6]. However, conditional independence of observations is usually assumed in the previous work, which is too restrictive for contextual modeling of visual scene. Compared to generative models including MRF and HMM, the conditional random field (CRF) relaxes the strong independence assumption and captures dependencies between observations [9]. In recent years, CRF has been applied to image labeling as well as video analysis [1,8,16,25,26].

Based on the CRF, this paper proposes a joint random field (JRF) model for visual scene modeling and presents its application to moving vehicle detection in grayscale video. The JRF model extends the CRF by introducing auxiliary latent variables to characterize complex visual environment and enhance moving object detection in video sequences, so that detection labels (e.g. vehicle/roadway) and hidden variables (e.g. intensity of shadowed points) are jointly estimated throughout the labeling process. A real-time algorithm of moving vehicle detection has been developed for video based traffic monitoring. The method handles both moving cast shadows/lights and dynamic background processes, and it integrates data-dependent contextual dependencies among both detection labels and hidden variables during the detection process. Experimental results show that the proposed approach effectively captures contextual information in video sequences and significantly improves the accuracy of moving vehicle detection under various illumination conditions.

## 2 Joint Random Field

Given an image sequence, the label and observation of a point  $x$  at time instant  $t$  are denoted by  $l_x^t$  and  $d_x^t$  respectively. The detection label  $l_x^t$  assigns the point  $x$  to one of  $K$  classes.  $l_x^t = e_k$  if the point  $x$  belongs to the  $k$ th class, where  $e_k$  is a  $K$ -dimensional unit vector with its  $k$ th component equal to one. The local observation  $d_x^t$  consists of intensity information at the site  $x$ . Here  $t \in \mathbf{N}$ ,  $x \in X$ , and  $X$  is the spatial domain of the video scene. The entire label field and observed image over the scene are compactly expressed as  $l^t$  and  $d^t$  respectively. Under complex visual environment, it is expected that image labeling can be enhanced by introducing a set of auxiliary latent variables to characterize the video scene over time. At time  $t$ , the hidden variable for each site  $x$  is denoted by  $r_x^t$ , and the entire latent field is expressed as  $r^t$ . In this work,  $K = 3$  and  $r_x^t$  describes the cast shadow/light at site  $x$  (see Section 3). Based on conditional random field, contextual information within both label field and latent field can be formulated through a probabilistic discriminative framework of statistical dependencies among neighboring sites.

### 2.1 JRF Model

For random variables  $\mathbf{v}$  and observed data  $\mathbf{o}$  over the video scene,  $(\mathbf{v}, \mathbf{o})$  is a conditional random field if, when conditioned on  $\mathbf{o}$ , the random field  $\mathbf{v}$  obeys the Markov property [9]:  $p(v_x | \mathbf{o}, v_y, y \neq x) = p(v_x | \mathbf{o}, v_y, y \in N_x)$ , where the set  $N_x$  denotes neighboring sites of the point  $x$ . Hence  $\mathbf{v}$  is a random field globally conditioned on the observed data. In order to introduce auxiliary hidden variables during the labeling process, the notion

of joint random field (JRF) is proposed in this work. For two random fields  $\mathbf{u}$ ,  $\mathbf{v}$  and observed data  $\mathbf{o}$ ,  $(\mathbf{u}, \mathbf{v}; \mathbf{o})$  becomes a joint random field if  $p(u_x, v_x | \mathbf{o}, u_y, v_y, y \neq x) = p(u_x, v_x | \mathbf{o}, u_y, v_y, y \in N_x)$ , i.e. the couple  $(\mathbf{u}, \mathbf{v})$  is Markovian when conditioned on observed data  $\mathbf{o}$ .

In this work, given the observed image  $\mathbf{d}^t$  at time instant  $t$ , the joint probability distribution over the label field  $\mathbf{l}^t$  and the latent field  $\mathbf{r}^t$  is modeled by a joint random field  $(\mathbf{l}^t, \mathbf{r}^t; \mathbf{d}^t)$  to formulate contextual dependencies. Thus the couple  $(\mathbf{l}^t, \mathbf{r}^t)$  obeys the Markov property when the observed data  $\mathbf{d}^t$  is given. Using the Hammersley-Clifford theorem and considering only up to pairwise clique potentials [22], the posterior probability is given by a Gibbs distribution with the following form.

$$p(\mathbf{l}^t, \mathbf{r}^t | \mathbf{d}^t) \propto \exp\left\{-\sum_{x \in X} [V_x(l_x^t, r_x^t | \mathbf{d}^t) + \sum_{y \in N_x} V_{x,y}(l_x^t, r_x^t, l_y^t, r_y^t | \mathbf{d}^t)]\right\}. \quad (1)$$

The one-pixel potential  $V_x(l_x^t, r_x^t | \mathbf{d}^t)$  reflects the local constraint for a single site. The two-pixel potential  $V_{x,y}(l_x^t, r_x^t, l_y^t, r_y^t | \mathbf{d}^t)$  imposes the pairwise constraint between neighboring sites. Strength of the constraints is dependent on the observed data. To simplify the computation, the pairwise potential is further factorized as  $V_{x,y}(l_x^t, l_y^t | \mathbf{d}^t) + V_{x,y}(r_x^t, r_y^t | \mathbf{d}^t)$ . Hence

$$p(\mathbf{l}^t, \mathbf{r}^t | \mathbf{d}^t) \propto \exp\left\{-\sum_{x \in X} [V_x(l_x^t, r_x^t | \mathbf{d}^t) + \sum_{y \in N_x} V_{x,y}(l_x^t, l_y^t | \mathbf{d}^t) + \sum_{y \in N_x} V_{x,y}(r_x^t, r_y^t | \mathbf{d}^t)]\right\}. \quad (2)$$

The JRF model extends the CRF for image sequences by introducing auxiliary latent variables to characterize complex visual scene, and it captures data-dependent neighborhood interaction among both detection labels and latent variables during the labeling process.

**Related work:** Recently, random field based models, such as hidden conditional random field and layout consistent random field, have been proposed to incorporate hidden variables for object/gesture recognition as well as segmentation of partially occluded objects [16,26]. Firstly, in these models and their extensions [5,13], labels are conditionally independent of observations given the hidden variables. In the proposed model, the observations impact the estimation of labels even when the hidden variables are given. In relatively complex visual processes such as moving vehicle and cast shadow detection, actually the detection labels (e.g. vehicle/roadway) are influenced by the observed images even when the hidden variables (e.g. pixel intensity under shadow) are known. From this point of view, the proposed model theoretically generalizes previous ones with a tradeoff in computational complexity. Hence the proposed model can be applied to gesture recognition and multi-object segmentation as well. Secondly, the auxiliary variables are continuous in this work, so that each site has a discrete detection label and a continuous hidden variable. Usually both the image label and the hidden variable of each site are discrete in the previous work. Thirdly, comparing to the proposed model, direct interaction (or constraint) between neighboring labels is ignored in previous approaches [13,26].

## 2.2 Optimization

The maximization of the joint posterior distribution over label field and latent field involves both discrete variables  $\mathbf{l}^t$  and continuous variables  $\mathbf{r}^t$ , which makes it difficult to

directly apply popular optimization methods for image labeling such as belief propagation and graph cut. The posterior probability is optimized by variational approximation [24]. The variational method looks for the best approximation of an intractable probability in the sense of Kullback-Leibler divergence. The posterior at time  $t$  can be approximated by the following distribution.

$$p(\mathbf{l}^t, \mathbf{r}^t | \mathbf{d}^t) \approx \prod_{x \in X} q(l_x^t | \mathbf{d}^t) q(r_x^t | \mathbf{d}^t), \quad (3)$$

where local approximating probabilities  $\{q(l_x^t | \mathbf{d}^t)\}$  and  $\{q(r_x^t | \mathbf{d}^t)\}$  are given by iterative computation. For each site  $x$  at time instant  $t$ ,  $\hat{l}_x^t = \arg \max_{l_x^t} q(l_x^t | \mathbf{d}^t)$  and  $\hat{r}_x^t = \arg \max_{r_x^t} q(r_x^t | \mathbf{d}^t)$ .

### 3 Moving Vehicle Detection

For video based traffic monitoring, each pixel in the scene is to be classified as moving vehicle, cast shadow (or light), or background (roadway). For a site  $x$  at time  $t$ , the label  $l_x^t$  equals  $e_1$  for background,  $e_2$  for shadow, and  $e_3$  for vehicle. Here static shadows are considered to be part of the background.

#### 3.1 Local Observation

In order to segment the moving vehicles, the system should first model the background and shadow information. For each point  $x$ , the pixel intensity  $d_x^t$  has three (R, G, and B) components for color images or one value for grayscale images. Grayscale images are considered in this work, while the formulation for color images can be derived similarly. Assume that each pixel in the scene is corrupted by Gaussian noise, so that the background model at time  $t$  becomes  $d_x^t = b_x^t + n_x^t$ , where  $b_x^t$  is the intensity mean for a pixel  $x$  within the background, and  $n_x^t$  is independent zero-mean Gaussian noise with variance  $(\sigma_x^t)^2$ . Intensity means and variances in the background can be estimated from previous images. For dynamic background scenes, the recent history of each pixel is modeled by a mixture of Gaussians during background updating [19]. As parameters of the mixture model change, the Gaussian distribution that has the highest ratio of weight over variance is chosen as the background model.

Given the intensity of a background point, a linear model is used to describe the change of intensity for the same point when shadowed (or illuminated) in the video scene, i.e.  $d_x^t = \rho_x^t b_x^t + n_x^t$ . Considering the contiguity of video image, the coefficient  $\rho_x^t$  can be estimated from its neighborhood as  $\rho_x^t \approx \frac{\sum_{y \in N_x} d_y^t}{\sum_{y \in N_x} b_y^t}$  if the point is under cast shadow.

To achieve maximum application independence, it is assumed that the intensity information of vehicles is unknown. Hence uniform distribution is used for the pixel intensity of moving vehicle. From the above discussion, the local intensity likelihood of a point  $x$  at time  $t$  becomes

$$p(d_x^t | l_x^t) = \begin{cases} N(d_x^t; b_x^t, (\sigma_x^t)^2), & \text{if } l_x^t = e_1 \\ N(d_x^t; \rho_x^t b_x^t, (\sigma_x^t)^2), & \text{if } l_x^t = e_2, \\ c, & \text{if } l_x^t = e_3 \end{cases} \quad (4)$$

where  $N(z; \mu, \sigma^2)$  is a Gaussian distribution with argument  $z$ , mean  $\mu$ , and variance  $\sigma^2$ ,  $c$  is a small positive constant ( $c = 1/256$  for grayscale images).

However, the observation model tends to confuse cast shadow and moving vehicle at boundary areas or in uniform regions, especially when the vehicle is darker than the background and the road surface is un-textured. Such detection error can be effectively reduced if the intensity of shadowed points is known, i.e.  $d_x^t = r_x^t + n_x^t$  if  $l_x^t = e_2$ , where  $r_x^t$  is the mean intensity under cast shadow (or light) for site  $x$ . Since the intensity under shadow is not given beforehand, in this work  $r_x^t$  is used as the auxiliary latent variable to characterize the visual scene for each point  $x$  at time  $t$ .

### 3.2 Contextual Constraint

The one-pixel potential in (2) is set as  $V_x(l_x^t, r_x^t | \mathbf{d}^t) = -\ln p(l_x^t, r_x^t | d_x^t)$ , so that posterior distribution  $p(\mathbf{l}^t, \mathbf{r}^t | \mathbf{d}^t)$  becomes the product of local posterior at each site  $\prod_{x \in X} p(l_x^t, r_x^t | d_x^t)$  when two-pixel potentials are ignored. Using the Bayes' rule,

$$p(l_x^t, r_x^t | d_x^t) \propto p(l_x^t, r_x^t, d_x^t) = p(l_x^t) p(d_x^t | l_x^t) p(r_x^t | l_x^t, d_x^t). \text{ Hence}$$

$$V_x(l_x^t, r_x^t | \mathbf{d}^t) = -\ln p(l_x^t) p(d_x^t | l_x^t) p(r_x^t | l_x^t, d_x^t). \quad (5)$$

The prior knowledge  $p(l_x^t)$  can be expressed by uniform distribution. The probability  $p(d_x^t | l_x^t)$  is given by the local intensity likelihood derived in the previous section. For pixel intensity under cast shadow (or light), the posterior  $p(r_x^t | l_x^t, d_x^t)$  is expressed as

$$p(r_x^t | l_x^t, d_x^t) = \begin{cases} N(r_x^t; \eta d_x^t + (1-\eta) \hat{r}_x^{t-1}, (\sigma_x^t)^2), & \text{if } l_x^t = e_2, \\ N(r_x^t; \hat{r}_x^{t-1}, (\sigma_x^t)^2), & \text{otherwise} \end{cases}, \quad (6)$$

where the positive  $\eta$  reflects the temporal continuity constraint for the auxiliary variable. Its value is set to be the same as the learning rate of background updating.

The two-pixel potential for neighboring detection labels is expressed as the following to formulate the spatial dependency.

$$V_{x,y}(l_x^t, l_y^t | \mathbf{d}^t) = -\alpha_1 l_x^t \cdot l_y^t - \frac{\alpha_2 l_x^t \cdot l_y^t}{\| (d_x^t - d_y^t) / \delta \|^2 + 1}, \quad (7)$$

where  $\delta^2 = \frac{1}{|N_x|} \sum_{i \in N_x} \|d_x^t - d_i^t\|^2 + \frac{1}{|N_y|} \sum_{i \in N_y} \|d_y^t - d_i^t\|^2$ , and  $\|\cdot\|$  is the Euclidean

distance. The first term (data-independent potential) encourages the formation of contiguous regions, while the second term (data-dependent potential) encourages data similarity when neighboring sites have the same label. The positives  $\alpha_1$  and  $\alpha_2$  respectively weight the importance of data-independent smoothness constraint and data-

dependent neighborhood interaction. However, under heavy noises neighboring sites may become quite different even though they belong to the same class. To prevent this problem when detecting vehicles within noisy video scene, the regulation term  $\delta$  is used in the data-dependent pairwise potential.

Similarly, the two-pixel potential for neighboring latent variables is expressed as

$$V_{x,y}(r_x^t, r_y^t | \mathbf{d}^t) = \beta_1 (r_x^t - r_y^t)^2 + \frac{\beta_2 (r_x^t - r_y^t)^2}{\| (b_x^t - b_y^t) / \varepsilon \|^2 + 1}, \quad (8)$$

where  $\varepsilon^2 = \frac{1}{|N_x|} \sum_{i \in N_x} \| b_x^t - b_i^t \|^2 + \frac{1}{|N_y|} \sum_{i \in N_y} \| b_y^t - b_i^t \|^2$ . The positives  $\beta_1$  and  $\beta_2$

respectively weight the importance of data-independent smoothness constraint and data-dependent neighborhood interaction for latent variables. The potential functions  $V_{x,y}(l_x^t, l_y^t | \mathbf{d}^t)$  and  $V_{x,y}(r_x^t, r_y^t | \mathbf{d}^t)$  capture neighborhood interactions among detection labels and latent variables respectively. Naturally, the potentials impose adaptive contextual constraints that will adjust the interaction strength according to the similarity between neighboring observations.

To balance the influence of potential terms for the joint random field, it is assumed that  $\alpha_1 = \alpha_2 = \alpha$  and  $\beta_1 = \beta_2 = \beta$ , where the parameters  $\alpha$  and  $\beta$  are empirically determined to reflect the constraint strength for detection labels and latent variables respectively. Initially,  $q(l_x^0) = \frac{1}{K}$  and  $q(r_x^0) = N(r_x^0; b_x^0, (\sigma_x^0)^2)$  with large variance for all the sites.

### 3.3 Preprocessing and Postprocessing

To improve the computational efficiency, the zone of moving vehicle detection is cropped from the scene for video processing (see Figure 1a). The region of interest is then straightened by applying perspective transformation [22], so that moving vehicle detection is performed on straightened images (see Figure 1b). The straightened image corresponds to a scaled top-down view of the roadway. Typically, a trapezoid region bounded by roadway lines becomes a rectangle with the prescribed width and length (48 by 72 in this work) in the straightened image. The image straightening reduces the number of pixels for subsequent video processing and substantially improves the computational efficiency. Bilinear interpolation is employed when warping the original image region onto the rectangle in the straightened image.

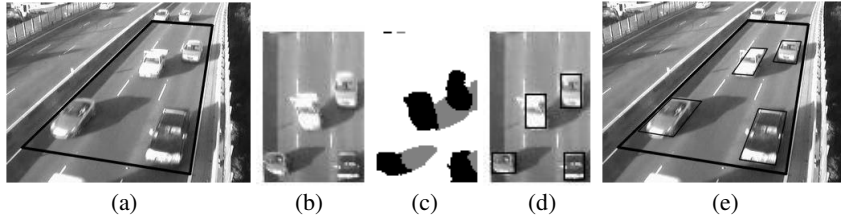


Figure 1. (a) Region of interest. (b) Straightened image. (c) Foreground and shadow detection. (d) Vehicle detection in the straightened image. (e) Vehicle detection in the original image.

After foreground detection with shadow removal (see Figure 1c), detected vehicles are approximated by small rectangles in the straightened image (see Figure 1d). For each detected foreground area, the corresponding rectangle has the same central point and average width and length, with its edges parallel to the horizontal and vertical axes. On the highway, most occlusions happen between moving vehicles from neighboring lanes. The detected roadway lines help separate occluded vehicles when occlusion across multiple lanes takes place. Small detected regions, such as frontal part of incoming vehicles and false detection caused by noises in the scene, are ignored to enhance the robustness of moving vehicle detection. The located rectangles in the straightened image are then mapped back onto the original image (see Figure 1e).

## 4 Results and Discussion

The proposed approach has been tested on grayscale video sequences captured under different environments for road traffic monitoring. The 48-pixel neighborhood is utilized in the algorithm. The C program can process about 25 frames per second on a Pentium 4 3.0G PC. Four moving vehicle and cast shadow detection algorithms are studied in our experiments: the mixture of Gaussians (MoG) based approach [10], the Markov random field (MRF) approach with spatiotemporal constraints [17], the dynamic conditional random field (CRF) approach [25], and the proposed joint random field (JRF) approach. The same initialization and neighborhood are used in these algorithms (when applicable).

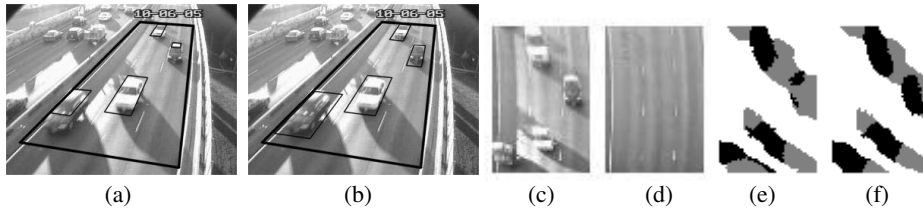


Figure 2. (a) Vehicle detection by the CRF approach. (b) Vehicle detection by the proposed approach. (c) Straightened image. (d) Estimated latent field (intensity of shadowed points). (e) Foreground and shadow detection by the CRF approach. (f) Foreground and shadow detection by the proposed approach.

Figure 2 shows the detection results by the conditional random field approach and the proposed method for a video sequence with strong reflection on the road surface. The gray regions in Figure 2e and 2f represent moving cast shadows. The CRF approach is unable to capture the intensity variation under relatively complex scene. It can be seen that parts of dark vehicles are misclassified in Figure 2e. On the other hand, the auxiliary variables (intensity of shadowed points) used in the proposed approach effectively model the illumination variation of the visual scene over time, which improves the reliability of vehicle and shadow segmentation. Compared with Figure 2e, moving vehicles and cast shadows at different locations of the road are accurately distinguished in Figure 2f.

Figure 3 shows the results of moving vehicle and cast shadow detection by the Gaussian mixture approach, the Markov random field approach, and the proposed method for a grayscale video sequence with low image contrast in the detection zone. In Figure 3e.1, the pixel based MoG approach is likely to confuse moving vehicle and cast

shadow under the noisy environment. The errors are corrected in Figure 3f.1 by the proposed method with the help of contextual dependencies and auxiliary variables (see Figure 3d.2). The MRF approach produces smooth segmentation results. However, sometimes it may smooth in a wrong way due to the neglect of the contextual interaction dependent on observations. It can be seen that some regions under shadow are misclassified in Figure 3e.2, while cast shadows are effectively removed from the moving vehicles in Figure 3f.2.

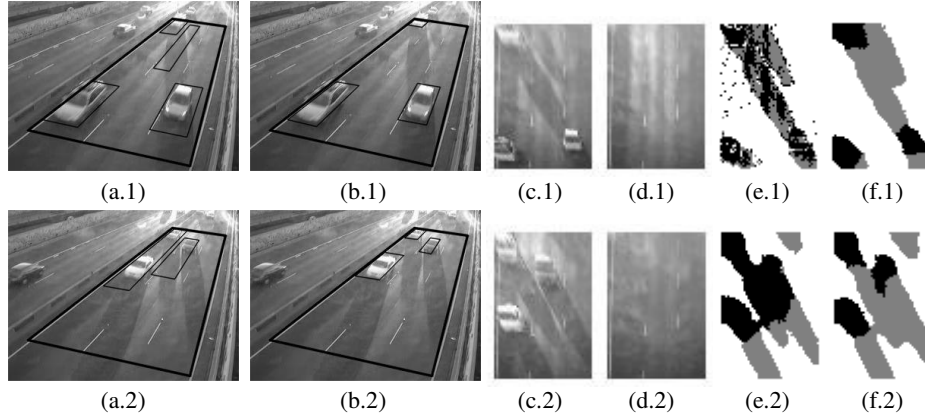


Figure 3. (a.1) Vehicle detection by the MoG approach. (a.2) Vehicle detection by the MRF approach. (b) Vehicle detection by the proposed approach. (c) Straightened images. (d.1) Estimated intensity of roadway. (d.2) Estimated latent field (intensity of shadowed points). (e.1) Foreground and shadow detection by the MoG approach. (e.2) Foreground and shadow detection by the MRF approach. (f) Foreground and shadow detection by the proposed approach.

The detection results are also evaluated quantitatively by comparing to the manually labeled ground-truth for straightened images. Table 1 shows the average error rate (portion of misclassified points in the entire image) for thirty representative frames of the three sequences (ten for each) shown in Figure 1-3. The MRF approach outperforms the MoG approach by utilizing smoothness constraints. Compared to the MRF approach, the CRF approach takes advantage of data-dependent neighborhood interactions. The JRF approach further improves the detection accuracy by introducing auxiliary latent variables to model the structure and evolution of the video scene. In our experiments, the JRF approach averagely reduces the error rate of the other three approaches by 73%, 58%, and 37% respectively. The substantial increase of the accuracy indicates that by integrating contextual constraints and introducing auxiliary variables, the proposed approach effectively models the traffic scene during the detection process.

	MoG	MRF	CRF	JRF
error rate	16.4%	10.2%	6.8%	4.3%

Table 1. Error rates of detection results.

Figure 4 shows the results of moving vehicle detection in the dark. The proposed approach can be applied to the detection of both cast shadows and cast lights. It can be seen that pixel intensity varies drastically when background points are illuminated by vehicle lights. Comparing to moving vehicles, the cast lights cover much more regions of the roadway, which could cause serious mistake and even failure in further video

analysis. The proposed method accurately discriminates cast lights from moving vehicles even in grayscale video sequences.

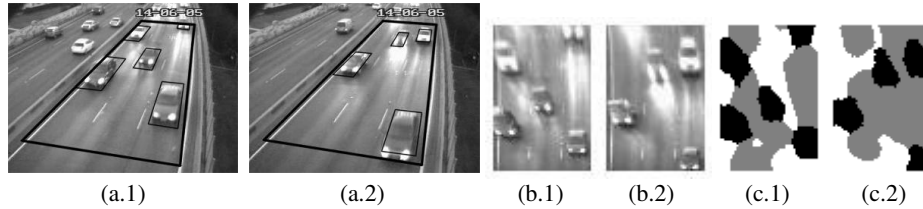


Figure 4. (a) Vehicle detection by the proposed approach. (b) Straightened images. (c) Foreground and light detection by the proposed approach.

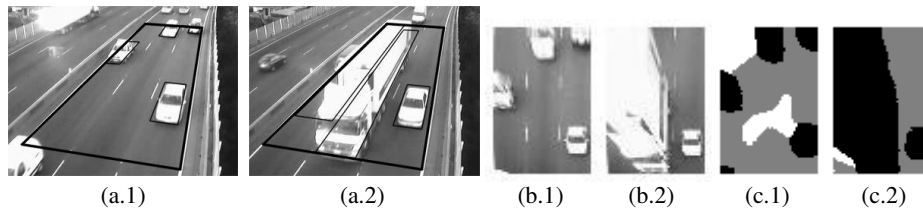


Figure 5. (a) Vehicle detection by the proposed approach. (b) Straightened images. (c) Foreground and shadow detection by the proposed approach.

Figure 5 shows the results of vehicle detection in video sequences with background illumination variations. The detected shadow regions are actually caused by the automatic gain control (AGC), since the setting has been automatically adjusted by the video camera when large vehicles pass by. Hence, even in video streams without moving cast shadows or lights, the proposed approach also helps prevent false vehicle detection under dynamic illumination variations.

## 5 Conclusion

There are two main contributions in this paper. The first is to propose a joint random field (JRF) model that extends CRF by introducing auxiliary latent variables to characterize visual scene over time and enhance moving object detection in video. The second is to develop a real-time vehicle detection algorithm for video based traffic monitoring. The proposed model integrates contextual constraints among both detection labels and hidden variables during the detection process. Experimental results show that the proposed approach effectively handles both cast shadows/lights and background illumination variations, and it significantly improves the performance of vehicle detection even in grayscale video sequences. Our future study is to apply the JRF model to activity/gesture recognition for video based event detection and develop traffic analysis techniques such as vehicle counting and tracking, stopped vehicle detection, and traffic flow estimation based on the proposed detection method.

## 6 Acknowledgement

National ICT Australia is funded through the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

## References

- [1] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, "Bilayer segmentation of live video," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 53-60, 2006.
- [2] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, pp. 1151-1163, 2002.
- [3] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," *Proc. Conf. Uncertainty in Artificial Intelligence*, pp. 175-181, 1997.
- [4] I. Haritaoglu, D. Harwood, and L. Davis, "W<sup>4</sup>: Real-time surveillance of people and their activities," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 22, pp. 809-830, 2000.
- [5] D. Hoiem, C. Rother, and J. Winn, "3D layoutCRF for multi-view object class recognition and segmentation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.
- [6] S. Kamijo, K. Ikeuchi, and M. Sakauchi, "Segmentations of spatio-temporal images by spatio-temporal Markov random field model," *Proc. EMMCVPR Workshop*, pp. 298-313, 2001.
- [7] J. Kato, T. Watanabe, S. Joga, J. Rittscher, and A. Blake, "An HMM-based segmentation method for traffic monitoring movies," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 24, pp. 1291-1296, 2002.
- [8] S. Kumar and M. Hebert, "Discriminative fields for modeling spatial dependencies in natural images," *Advances in Neural Information Processing Systems*, pp. 1351-1358, 2004.
- [9] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proc. Int'l Conf. Machine Learning*, pp. 282-289, 2001.
- [10] N. Martel-Brisson and A. Zaccarin, "Moving cast shadow detection from a Gaussian mixture shadow model," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 643-648, 2005.
- [11] I. Mikic, P. Cosman, G. Kogut, and M. Trivedi, "Moving shadow and object detection in traffic scenes," *Proc. Int'l Conf. Pattern Recognition*, vol. 1, pp. 321-324, 2000.
- [12] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 302-309, 2004.
- [13] L.-P. Morency, A. Quattoni, T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.
- [14] N. Paragios and V. Ramesh, "A MRF-based approach for real-time subway monitoring," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 1034-1040, 2001.
- [15] A. Prati, I. Mikic, M. Trivedi, and R. Cucchiara, "Detecting moving shadows: Algorithms and evaluation," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 25, pp. 918-923, 2003.
- [16] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 29, pp. 1848-1853, 2007.
- [17] J. Rittscher, J. Kato, S. Joga, and A. Blake, "A probabilistic background model for tracking," *Proc. European Conf. Computer Vision*, vol. 2, pp. 336-350, 2000.
- [18] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 27, pp. 1778-1792, 2005.
- [19] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 22, pp. 747-757, 2000.
- [20] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J. Buhmann, "Topology free hidden Markov models: Application to background modeling," *Proc. Int'l Conf. Computer Vision*, pp. 294-301, 2001.
- [21] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, "Background cut," *Proc. European Conf. Computer Vision*, vol. 2, pp. 628-641, 2006.
- [22] A. M. Tekalp, *Digital Video Processing*, Prentice Hall, 1995.
- [23] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," *Proc. Int'l Conf. Computer Vision*, vol. 1, pp. 255-261, 1999.
- [24] M. Wainwright and M. Jordan, "Graphical models, exponential families, and variational inference," Technical Report, University of California, Berkeley, 2003.
- [25] Y. Wang, K.-F. Loe, and J.-K. Wu, "A dynamic conditional random field model for foreground and shadow segmentation," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 28, pp. 279-289, 2006.
- [26] J. Winn and J. Shotton, "The layout consistent random field for recognizing and segmenting partially occluded objects," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 37-44, 2006.