

Bayesian Decision Fusion for Dynamic Multi-cue Object Detection*

Lucas Paletta and Gerhard Paar

Institute of Digital Image Processing, Joanneum Research

Wastiangasse 6, A-8010 Graz, Austria

{lucas.paletta,gerhard.paar}@joanneum.at

Abstract

Visual object detection using single cue information has been successfully applied in various tasks, in particular for near range recognition. While robust classification and probabilistic representation enhance 2D pattern recognition performance, they are 'per se' restricted due to the limited information content of single cues. The contribution of this work is to demonstrate performance improvement using multi-cue information integrated within a probabilistic framework. 2D and 3D visual information naturally complement one another, each information source providing evidence for the occurrence of the object of interest. We demonstrate preliminary work describing Bayesian decision fusion for object detection and illustrate the method by robust detection of traffic infrastructure, and add an extension to integrate belief dynamically from image sequences.

1 Introduction

Object recognition and detection based on visual information has been successfully applied in various tasks [9, 8, 26, 25], in particular for near range recognition [25, 12, 10, 20, 18]. Specific tasks impose additional challenges on the robustness of a detection system, such as outdoor imaging (e.g., illumination variations) or automatic object detection from preprocessed regions of interest (ROIs) in real-world images. To overcome these problems, robust recognition [10], illumination tolerant classification [2] and probabilistic detection [12, 20, 18] have been introduced to enhance the performance of 2D pattern recognition methods. However, performance gains from these methods remain restricted as long as they rely on the limited information content of single information cues.

The original contribution of this work is to demonstrate that the *integration of multi-cue visual information* improves detection performance within a *probabilistic frame-*

work. The essential role of information fusion in image understanding [23] and pattern recognition has already been sufficiently outlined. Though, most work on fusion focuses either on the integration of multi-source data [6] or on the dynamic accumulation of evidence from single-cue information [3, 19]. The utility of multi-cue evidence has been stressed for tracking issues [5] and visual servoing tasks [24]. The presented work outlines integration within the mathematical framework of Bayesian decision fusion and with respect to the context of visual object detection. Detection is here triggered by the fusion of 2D and 3D information which naturally complement one another, each information source providing evidence for the occurrence of the object of interest.

Multi-cue object detection is evaluated within experiments of a characteristic Mobile Mapping application. Mobile Mapping of environment information from a moving platform plays an important role in the automatic acquisition of GIS (Geographic Information Systems). The extraction of traffic infrastructure from video frames captured on a moving vehicle requires a robust visual object detection system that provides both high localization accuracy and the capability to cope with uncertain information [18]. The efficient extraction of vertical object 3D structure [14] and the robust detection of traffic signs using 2D appearance based object recognition [17] are now combined to give an improved estimate on the object identity and location within the video frame.

The method on probabilistic multi-cue information fusion is sketched as follows (Figure 1),

1. Object specific 3D reconstruction and range segmentation.
2. Probabilistic modeling of object relevant 3D information.
3. View based object detection using a probabilistic neural network.
4. Bayesian decision fusion of 2D and 3D multi-cue confidence support maps.
5. Maximum-A-Posteriori (MAP) classification with respect to the object confidence maps.

* This work is funded by the European Commission's IST project DETECT under grant number IST-2001-32157. This work has been carried out within the K plus Competence Center ADVANCED COMPUTER VISION. This work was funded from the K plus Program.

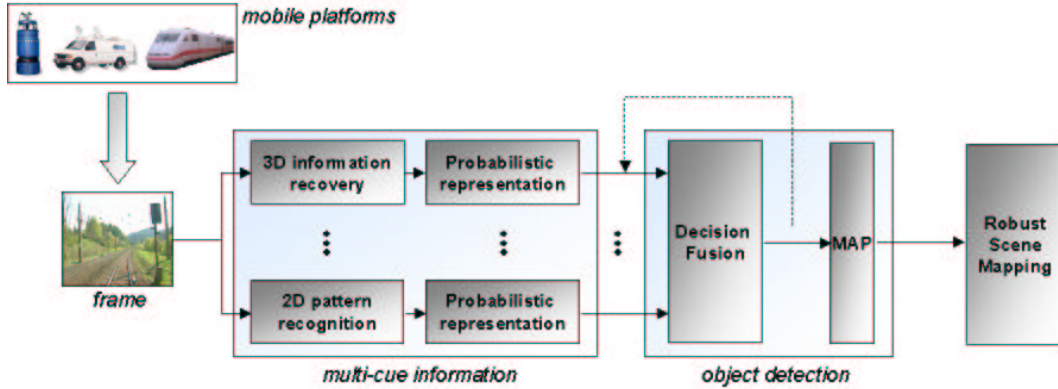


Figure 1: Concept of the object detection system using multi-cue information fusion.

The paper gives an outline of the probabilistic multi-cue object detection methodology and demonstrates preliminary results.

2 Probabilistic object localization from 3D information

In order to achieve a probabilistic representation of object location, the 3D information is first recovered from a video frame sequence. In Mobile Mapping applications, object location refers in many cases to a ground plane (road, railroad embankment, etc.). Redundant data on object height is therefore used for aggregation of object evidence which is here formulated within a probabilistic framework to enable segmentation and multi-cue fusion in the sequel.

2.1 Recovery of 3D information

3D reconstruction of the environment is here accomplished by structure from motion. Corresponding points in successive images are obtained by a stereo matching tool (*Hierarchical Feature Vector Matching, HFVM*, [15]) which has been adopted for the case of motion stereo [16]. It generates a dense disparity map (correspondences on almost each pixel). For 3D reconstruction, the orientation of the camera with respect to the moving vehicle [27] is determined in a calibration step. We assume odometry and velocity information to be available for each image. This enables, together with the system calibration, the exact orientation of each camera position with respect to the route and to determine both the distance to a matched point and the exact position within 3D space.

The idea of 3D object specific segmentation is based on the fact that - for many cases in Mobile Mapping - objects of interest are mounted vertical (Figure 2a,b). As a

consequence, the projection of all measured object points generates an aggregation on the horizontal plane (Figure 2c). Stored in a digital elevation model (DEM), these aggregations can be easily segmented, e.g., by lowpass filtering and thresholding. Backprojection of the identified segments gains ROI's in the input frame (Figure 2d). As a byproduct, for each pixel on these segments the distance as well as the global coordinates give important scaling information for the following object recognition steps. Additional valuable information such as a prediction for the track angles in the image, a prediction for the sky region, or the Focus of Expansion (FOE) can be extracted directly from the orientation data.

2.2 Probabilistic representation of object location

Each single object location - which has been derived from a point aggregation (Section 2.1) - implicitly represents uncertain information. We propose to model this local uncertainty by a multivariate unimodal Gaussian $\varphi_j(\mathbf{y})$,

$$\varphi_j(\mathbf{y}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\},$$

with mean $\boldsymbol{\mu}_j$ and covariance matrix Σ_j and with respect to a sample \mathbf{y} within the ground plane. $\varphi_j(\mathbf{y})$ represents thus the probability density function given an object o_j by $p(\mathbf{y}|o_j)$ (Figure 3(a)).

For each video frame and its mapping of 3D locations onto the ground plane, one can automatically find the appropriate locations of means, $\boldsymbol{\mu}_j$, by applying a clustering scheme. A statistically efficient useful cluster algorithm which naturally makes sense out of these local Gaussian distributions, is the *expectation-maximization* (EM) algorithm [7]. It approximates an entire distribution of samples by a

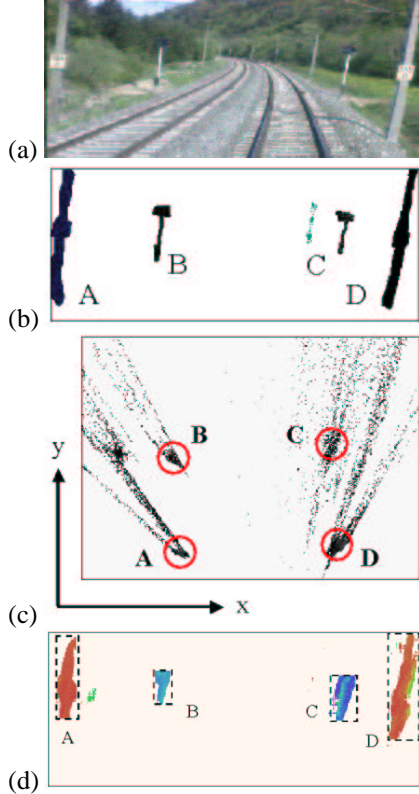


Figure 2: Object specific segmentation of 3D information. (a) Video frame of reference, (b) vertically accentuated 3D structure (A-D), (c) 3D point aggregations from motion stereo, (d) associated 2D regions of interest.

mixture density model, i.e.,

$$p(\mathbf{y}) = \sum_{j=1}^M P(j)\varphi_j(\mathbf{y}) \quad (1)$$

where the parameters $P(j)$ are the mixing coefficients. $P(j)$ can be regarded as prior probabilities for the data points to have been generated from the j th component of the mixture. EM iteratively determines appropriate means and covariances so as to maximize the likelihood of the data with respect to this model.

Each single cluster kernel - represented by the Gaussian - is then assumed to represent the localization uncertainty with respect to a single local 3D object. These confidence values are then backprojected into the input frame according to Section 2.1 (Figure 3, 6(b)). Backprojected points are members of cluster j (up to some confidence threshold) and result in a confidence support map with respect to object specific 3D information.

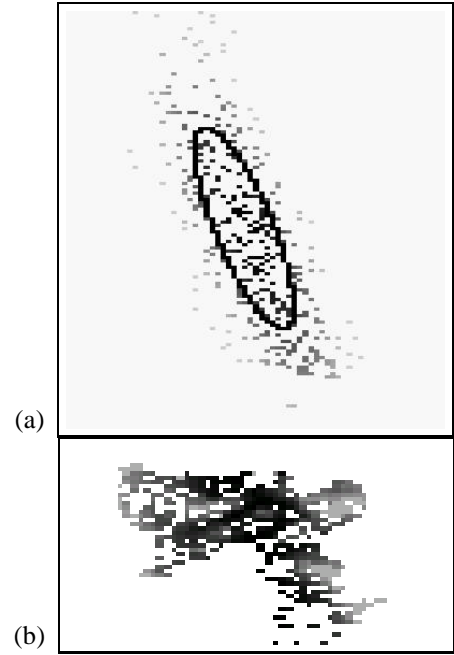


Figure 3: (a) Single-class Gaussian with ellipsoid of uniform Mahalanobis distance to mean μ_j superimposed, (b) projected confidences into 3D object related ROIs (zoomed out from Figure 6(b)).

3 Probabilistic view based object detection

Object recognition based on 2D information is a further operation concerned in a multi-cue detection scheme. The classification is based on a model database of image templates which were, e.g., manually segmented from real imagery. Efficient object localization and detection is correspondingly outlined in [12, 17]. The presented work outlines appearance based pattern matching in a probabilistic framework [12, 21, 19] to quantify the level of uncertainty in the classification and hence further enable reasoning on the dynamics of visual information.

Appearance based representation The detection process is based on a recognition module operating on local image patterns which are successively extracted from the image (Figure 4). Appearance based object representations [13] consist of a collection of raw sensor footprints combining effects of shape and reflectance [12, 21, 19]. In contrast, geometric models suffer from matching complexity and fail to work for complex shapes [8]. Instead of storing high-dimensional pixel patterns \mathbf{x} , the sensor vector can be transformed by principal component analysis (PCA) to a low-dimensional representation \mathbf{y} in feature space, called *eigenspace* [13]. It captures the maximum variations in the

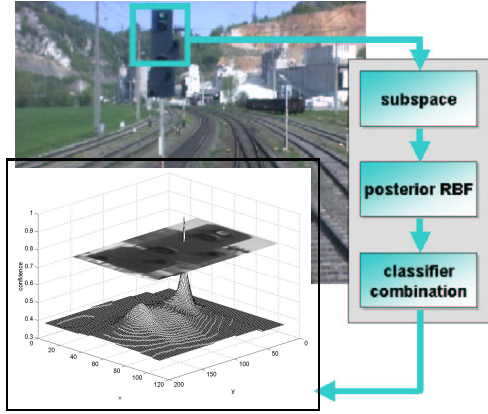


Figure 4: Object detection of traffic signs. Subwindows from the image are projected to eigenspace (PCA) and mapped by RBF networks for a probabilistic interpretation.

presented data set whereas distances are a measure of image correlation [13, 12]. Recognition is supported by the property that close points in subspace correspond to similar object appearances.

Probabilistic matching Object representations with models of uncertainty in eigenspace require estimates of the data density [12]. The present system uses this concept under definition of a *rejection class* w.r.t. background for a closed world interpretation [21]. A posterior neural classifier maps then the PCA description to a distribution over predefined object classes [21, 18]. Radial basis functions (RBF) networks [4, 21] apply a Bayesian framework with density estimations provided by unsupervised clustering, where the confidence estimates are refined by supervised learning. The feature vector \mathbf{y} is fed to the network and mapped to the output z_κ , $\kappa = 1.. \Omega$, Ω is the number of objects, for a posterior estimate, $\hat{P}(o_\kappa|\mathbf{y}) = \alpha z_\kappa(\mathbf{y})$, α is a normalizing constant. A decision on object recognition is applied using a Maximum A Posteriori (MAP) decision on z_κ .

4 Multi-cue decision fusion for object detection

Bayesian decision fusion Fusion with respect to 2D and 3D information on object specific evidence is here applied to the corresponding posterior estimation, i.e., the belief distributions related to 2D and 3D information. In particular, Bayesian decision fusion [1, 6] is operated on the 2D and 3D multi-cue confidence support maps: A naive Bayes classifier [17] represents then the simplified Bayesian update of the probability distribution on object hypotheses (results in Figure 6(d)).

The fusion method is outlined as follows. In a set of $\gamma = 1.. \Gamma$ different confidence support maps, global confidence in the classification is updated by *fusion* of a 'current' cue specific belief $\hat{p}(o_\kappa|\mathbf{g}_\gamma)$ with the integrated hypotheses $\hat{p}(o_\kappa|\mathbf{y}_1, \dots, \mathbf{y}_{\Gamma-1})$. The overall belief in hypothesis o_κ is calculated by Bayesian inversion [22], $\hat{p}(o_\kappa|\mathbf{y}_1, \dots, \mathbf{y}_\Gamma) = \alpha \hat{p}(\mathbf{y}_1, \dots, \mathbf{y}_\Gamma|o_\kappa) \hat{p}(o_\kappa)$, where α is a normalizing constant. Recursive updating is simplified assuming conditional independence of the measurements [22] which implies

$$\hat{p}(o_\kappa|\mathbf{y}_1, \dots, \mathbf{y}_\Gamma) = \alpha \hat{p}(o_\kappa) \prod_{\gamma=1}^{\Gamma} \hat{p}(\mathbf{y}_\gamma|o_\kappa). \quad (2)$$

A local decision on object identity is then performed via Maximum-A-Posteriori (MAP) [11] classification with respect to a location represented in the Γ confidence maps.

Dynamic fusion from image sequences We consider several methods to improve the multi-cue fusion method using *information integration over time*. Firstly, we concern the clustering of probabilistic estimation of 3D location (Section 2) and apply a motion model using a Kalman filter based tracking approach. This results in a trajectory of probabilistic object location over time. Secondly, the tracked 2D information can be integrated with Bayesian decision fusion, according to $\hat{p}(o_\kappa|\mathbf{y}_{\Gamma,t}, \mathbf{y}_{\Gamma,t-1}) = \beta \hat{p}(\mathbf{y}_{\Gamma,t}, \mathbf{y}_{\Gamma,t-1}|o_\kappa) \hat{p}(o_\kappa)$.

5 Experimental results

The presented multi-cue detection system is a general purpose system to automatically localize objects such as traffic signs [17], subway or railway objects [18], etc. The images used for the experiment were captured from top of the *measurement waggon* of the Austrian Federal Railways, during a regular train trip from Vienna to Graz.

For the 2D detection classifier, the posterior belief function was estimated by a radial basis functions (RBF) neural network classifier which was trained using 724 sample templates from 7 highly relevant sign classes. The evidence contributed by different R,G,B channels was fused according to a classifier combination [17] to receive increased detection performance, i.e., $\approx 89\%$ recognition accuracy on the complete test set, including severe illumination changes and noise in the image extraction [18]. A detailed description of the 2D recognition experiments is found in [17].

The performance of the 3D segmentation method was monitored on extended video frame sequences, mostly demonstrating robust performance [16, 18]. However, in rare cases the 3D information was not recovered, possibly due to the large extent of visual motion which is encountered when the observer is in the process of passing by.

class	symbol	sample
<i>Hauptsignal-HS</i>		
<i>Hauptsignal (back)-HSb</i>		
<i>Vorsignal-VS</i>		
<i>Vorsignal (back) -VSb</i>		
<i>Fahrleitungssignal-FS</i>		
<i>Geschwindigkeitstafel-GT</i>		
<i>Signalnachahmer-SNA</i>		

Table 1: Object classes for traffic light/sign recognition (object terminology according to Austrian Federal Railways).

Since a detection system must minimize its resulting *negative false* classifications and should not overlook any objects along the route, these cases require even more robust methods as the presented multi-cue information fusion.

Figure 6(a) depicts a typical video frame from a railway route including a near range object (traffic light). Here, the resulting scatter image of the ground plane (Figure 5(a)) will not enable an accurate localization. Therefore, the scatter image is processed by the EM clustering algorithm (Section 2.2, Figure 5(b)) to provide a probabilistic representation of object location. The cluster points are then back-projected into 2D (Figure 6(b)) to enable information fusion (section 4). Figure 6(c) illustrates the confidence support map as result of the 2D classifier. The final confidence map according to pixel-wise multi-cue decision fusion is presented in Figure 6(d). It is clearly seen that the fusion operation is capable to 'wash out' multiple erroneous and ambiguous confidence values from 3D and 2D processing.

6 Discussion

The presented work provides a system prototype that successfully demonstrates the concept of multi-cue - i.e., 2D and 3D - information fusion within a probabilistic framework, with the aim to render object detection more robust. The method represents a starting point for more complex Mobile Mapping systems that would be capable to perform

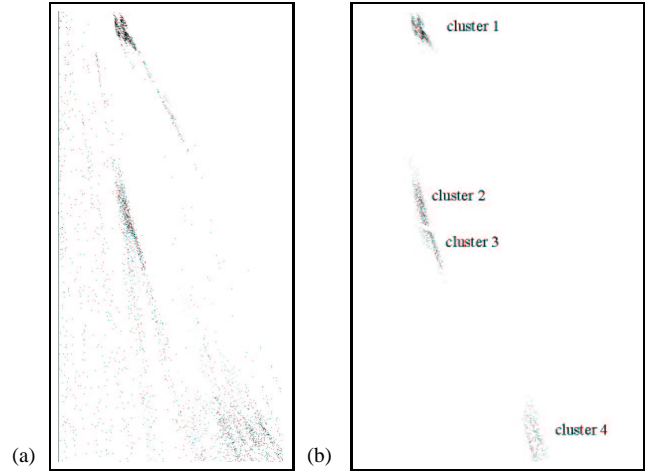


Figure 5: (a) Scatter image with origin of observer to object distance (x,y) at the top left corner, (b) location specific probability distributions extracted by EM clustering algorithm (Section 2.2).

reasoning for the efficient use of uncertain multi-cue visual information.

This paper demonstrates preliminary work which we account as a promising basis to profoundly investigate multi-cue fusion with respect to various informatio sources. Future work will focus on extended statistical evaluations of the presented system, the effect on multi-frame tracking and decision fusion on spatio-temporal cues, and on attention based mechanisms that enable efficient use of the given visual information.

References

- [1] M. A. Abidi and R. C. Gonzalez, editors. *Data Fusion in Robotics and Machine Intelligence*. Academic Press, San Diego, CA, 1992.
- [2] H. Bischof, H. Wildenauer, and A. Leonardis. Illumination insensitive eigenspaces. In *Proc. ICCV01*, volume 1, pages 233–238. IEEE Computer Society, 2001.
- [3] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. Appearance-based active object recognition. *Image and Vision Computing*, 18(9):715–727, 2000.
- [4] D. S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.
- [5] J. L. Crowley and F. Berard. Multi-modal tracking of faces for video communications. In *Proc. Conference on Computer Vision and Pattern Recognition*, 1997.
- [6] B. Dasarthy. *Decision Fusion*. IEEE Computer Society Press, Los Alamitos, CA, 1994.
- [7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38, 1977.
- [8] S. Edelman. Computational theories of object recognition. *Trends in Cognitive Sciences*, 1:296–304, 1997.
- [9] S. Grossberg, H. Hawkins, and A. Waxman. Special Issue - Automatic Target Recognition. *Neural Networks*, 8:1003–1360, 1995.

- [10] A. Leonardis and H. Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding*, 78(1):99–118, 2000.
- [11] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, NY, 1997.
- [12] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.
- [13] H. Murase and S. K. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.
- [14] G. Paar. Segmentation of vertical objects using motion-based stereo. In *Proc. Workshop of the AAPR*, pages 49–54. Berchtesgaden, Germany, 2001.
- [15] G. Paar and W. Pölzleitner. Robust disparity estimation in terrain modeling for spacecraft navigation. In *Proc. International Conference on Pattern Recognition*, The Hague, Netherlands, 1992.
- [16] G. Paar, O. Sidla, and W. Pölzleitner. Genetic feature selection for highly-accurate stereo reconstruction of natural surfaces. In *Proc. SPIE Conference on Intelligent Robots and Computer Vision XVII, Paper 3522-50*, 1998.
- [17] L. Paletta. Detection of railway signs using posterior classifier combination. In *Proc. International Conference on Pattern Recognition*, Quebec City, Canada, 2002, *in print*.
- [18] L. Paletta, G. Paar, and A. Wimmer. Mobile visual detection of traffic infrastructure. In *Proc. IEEE International Conference on Intelligent Transportation Systems*, pages 616–621, Oakland, CA, 2001.
- [19] L. Paletta, M. Prantl, and A. Pinz. Learning temporal context in active object recognition using Bayesian analysis. In *Proc. International Conference on Pattern Recognition*, pages 695–699, 2000.
- [20] L. Paletta and E. Rome. Learning fusion strategies for active object detection. In *Proc. International Conference on Intelligent Robots and Systems*, pages 1446–1452. Takamatsu, Japan, 2000.
- [21] L. Paletta, E. Rome, and A. Pinz. Visual object detection for autonomous sewer robots. In *Proc. International Conference on Intelligent Robots and Systems, IROS'99*, pages 1087–1093. Kyongju, South Korea, 1999.
- [22] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988.
- [23] A. Pinz and R. Bartl. Information fusion in image understanding. In *Proc. International Conference on Pattern Recognition*, pages 366–370. Silver Springs, MD, 1992.
- [24] P. Pirjanian, J.A. Fayman, and H. I. Christensen. Improving task reliability by fusion of redundant homogeneous modules using voting schemes. In *Proc. IEEE International Conference on Robotics and Automation*, pages 425–430, 1997.
- [25] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 317:314–319, 1990.
- [26] F. Sadjadi. *Automatic Target Recognition XII*. Proc. of SPIE Vol. 4726, Aerosense 2002, Orlando, FL, 2002.
- [27] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Transactions on Robotics and Automation*, 3(4):323–344, 1987.

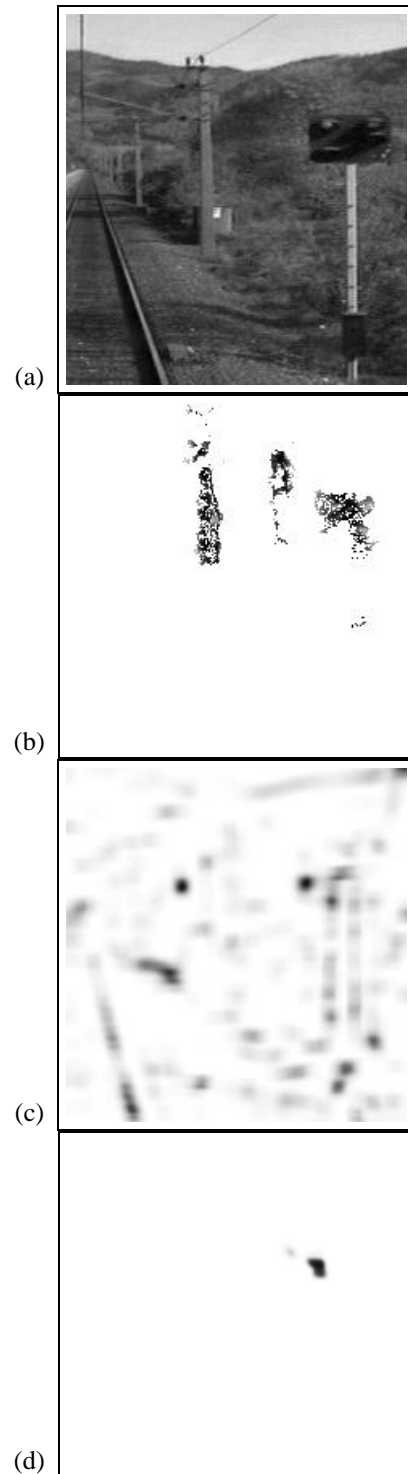


Figure 6: (a) Original image, (b) confidence ROIs from original image - high confidences in black, (c) confidence results from scanned 2D information object interpretation, (d) confidence map fused from 2D and 3D information.