Uncertain Geometry: A New Approach to Modeling for Recognition

Joseph L. Mundy and Ozge C. Ozcanli
Brown University, Box D, Providence, Rhode Island USA

ABSTRACT

Over the last several years, a new representation for geometry has been developed, based on a 3-d probability distribution of surface position and appearance. This representation can be constructed from multiple images, using both still and video data. The probability for 3-d surface position is estimated in an on-line algorithm using Bayesian inference. The probability of a point belonging to a surface is updated as to its success in accounting for the intensity of the current image at the projected image location of the point. A Gaussian mixture is used to model image appearance. This update process can be proved to converge under relatively general conditions that are consistent with aerial imagery. There are no explicit surfaces extracted, but only discrete surface probabilities. This paper describes the application of this representation to object recognition, based on Bayesian compositional hierarchies.

Keywords: Object recognition, 3-d model, Bayesian inference

1. INTRODUCTION

The performance of object recognition algorithms in visual imagery is limited by many effects such as varying viewpoint, varying illumination and occlusion. In addition, the segmentation of an object from background can be very difficult given complex scene content. This paper introduces a new approach to overcome these difficulties through a new probabilistic representation of the 3-d world based on volume elements (voxels). Each volumetric element of space contains a probability that an object surface exists at that location and a statistical model for the appearance of the hypothesized surface. These models provide a measure of the “normal” appearance of the world. This framework can thus detect deviations from normalcy, i.e. change detection. The detected changes are then regions of interest to be further characterized by a compositional object recognition algorithm that operates in the same probabilistic framework. As will be described, this framework provides important context for recognition including: prediction of occlusion; prediction of shadows; and the appearance of background surfaces.

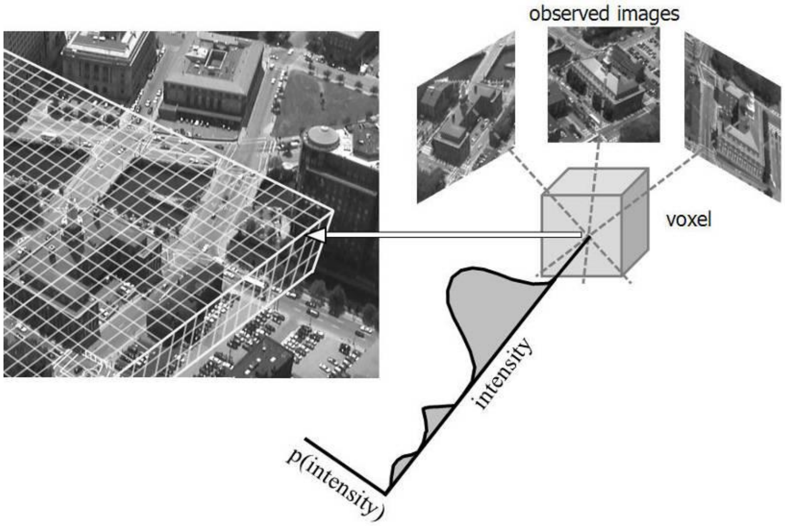


Figure 1 The volumetric appearance model concept. Three dimensional space is decomposed into cubical regions called voxels. Each voxel contains a Gaussian mixture distribution to measure the probability density of observed image intensity or color. A camera model is available for each image so that image pixels are projected into the voxel grid along rays. This intensity information updates the appearance distribution and the existence of 3-d surfaces is hypothesized using a Bayesian inference algorithm.

The representation is shown in Figure 1. A region of 3-d space is delineated by a volumetric model consisting of elements called *voxels* (volume elements). Ideally, a voxel is either empty or contains a surface element that bounds a solid region of space. The surface voxels are associated with a model of appearance as observed in images. In this paper, the image modalities are restricted to visible wavelengths, either grayscale or color. In general, the model can handle non-literal imagery but that possibility will not be considered here. The model for image appearance is a Gaussian mixture distribution that can account for a range of variability due to illumination direction, shadows and image misregistration. It is also assumed in this work that a camera model (geometric projection from 3-d to 2-d) is supplied with each image to be processed. In general it is necessary to calibrate and register such projection models due to errors in platform motion and internal parameters, but these adjustments are outside the scope of this paper. However, it should be mentioned that the voxel framework is capable of supporting automated camera calibration and registration.

It is not assumed that there is a fixed set of images available at the beginning and the model of the world is built up incrementally as images are observed in an unbounded sequence. This type of processing is called an *on-line* algorithm in computer science terminology. It is generally not known a priori what voxels contain surface elements and so the appearance models are stored at each volume element and adapted as each image is processed. The probability that a voxel, X , contains a surface element, $P(X \in S)$, is updated concurrently. Thus, the process is one of joint estimation where 3-d surface geometry and appearance are learned from an image sequence. The joint estimation process takes into account the success of the appearance models in accounting for the observed image intensity as well as how likely a voxel is to contain the observed surface, given the possibility of occlusion. Thus, each pixel in a new image can be assigned a probability as to how well the observed intensity or color agrees with those values observed in past images. This analysis provides a mechanism for *change detection* where pixel colors or intensity that have low probability represent areas that have changed with respect to the condition of the world observed in the past.

2. LEARNING GEOMETRY AND APPEARANCE

2.1 The Gaussian mixture appearance model

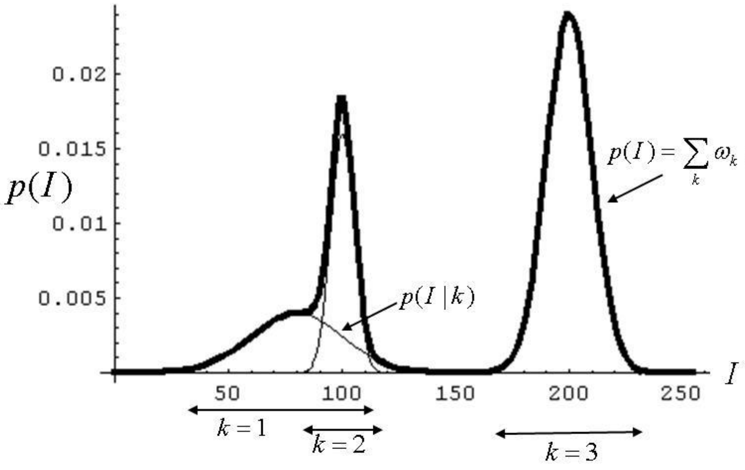


Figure 2 The Gaussian mixture appearance model. The probability density for observed intensity can be a mix of different states such as shadow or occlusion.

Each voxel contains a Gaussian mixture distribution as illustrated in Figure 2. For this particular example there are three mixture components, $p(I \mid k)$ and associated mixing weights, ω_k. The form of $p(I \mid k)$ is given by,

$$p(I \mid k) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\Sigma_k\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(I-\mu_k)^{\,t}\,\Sigma_k^{-1}(I-\mu_k)},$$

where the pixel observation, I, can be considered a vector with one or more elements to account for color. The quantities, μ_k and Σ_k, are the mean and covariance of the Gaussian distribution. The mixture itself is just a weighted sum of some number of these components.

$$p(I) = \sum \omega_k p(I \mid k)$$

The parameters of the mixture are learned using a modified expectation maximization (EM) algorithm similar to that used in video background modeling[1]. The update of the parameters is as follows.

$$\omega_k^{N+1} = \omega_k^N + d\omega_k^{N+1}$$

$$\mu_k^{N+1} = \mu_k^N + \frac{d\omega_k^{N+1}}{\omega_k^N + d\omega_k^{N+1}} \left(I^{N+1} - \mu_k^N \right)$$

$$\Sigma_k^{N+1} = \Sigma_k^N + \frac{d\omega_k^{N+1}}{\omega_k^N + d\omega_k^{N+1}} \left[\left(I - \mu_k^N \right)^t \left(\Sigma_k^N \right)^{-1} \left(I - \mu_k^N \right) - \Sigma_k^N \right]$$

The increment in mixing weight, $d\omega_k^{N+1}$, upon observing image $N+1$ is determined by analyzing the distributions in other voxels along the same camera ray, that could contribute to pixel intensity, I. This computation is described in the next section. The number of components of the distribution is adapted as necessary to account for image intensities (colors) as new images are observed. If a narrow range of intensity values are observed over the image sequence, then the mixing probability of the nearest component will approach one and the density will be sharply peaked around the mean. Only mixture components with means within a few standard deviations of the observed color are updated as described by the equations above.

2.2 Surface probability and occlusion

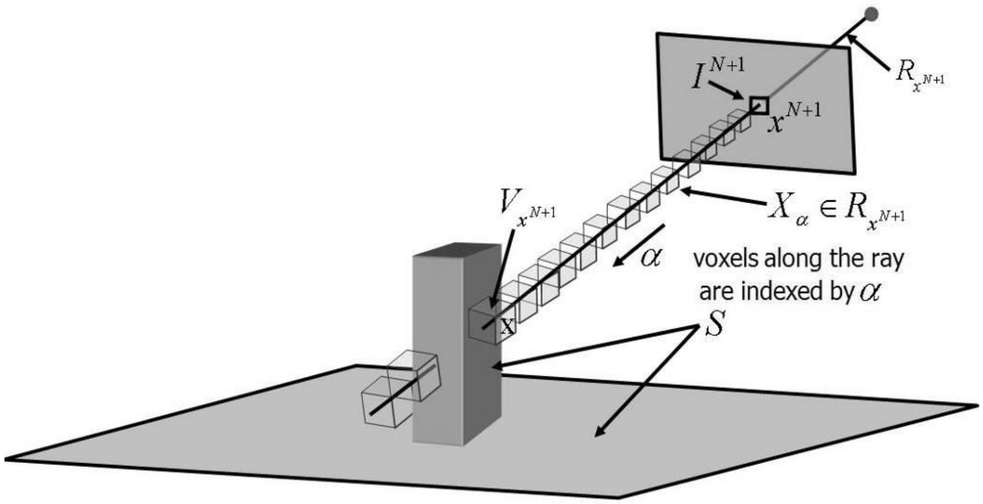


Figure 3 In order to properly update voxel appearance models, it is necessary to consider occlusion by multiple surfaces along a ray. Shown here is a ray through image pixel location, x^{N+1}, for image $N+1$. The content of each voxel has been updated for all N previous images. There are three possible surfaces that could explain the observed image intensity, I^{N+1}. The actual surface that produced the intensity is contained in voxel X.

In order to determine the update to the mixture distribution of a particular voxel it is necessary to consider all the voxels along the same ray through the voxel and the corresponding image pixel location as shown in Figure 3. In the figure a ray, $R_{x^{N+1}}$, is cast through an image pixel at location x^{N+1}in image $\Im^{N+1}$. This ray is shown to intersect several surfaces in the world. The location of voxels that contain surfaces are not known with certainty, and so each voxel has some probability of containing the surface that gives rise to the observed image intensity, I_x^{N+1}, at image location x^{N+1}. At the same time, if there are voxels closer to the camera, X_α, with high surface probabilities then they will occlude the voxels further along the ray and make it less likely that a voxel such as X could have produced the observed image intensity. The voxel that actually gives rise to the observed image intensity is denoted by$V_{x^{N+1}}$. The posterior probability for the existence of a surface element in voxel X is given by,

$$P^{N+1}(X \in S) = P^N(X \in S)\frac{p^N(I_x^{N+1} \mid X \in S)}{p^N(I_x^{N+1})}$$

This update relation has a relatively intuitive interpretation - the probability of a voxel X containing a surface is increased if the mixture distribution in that voxel assigns the observed intensity a higher probability density than any other voxel along the ray. The responsible voxel must also not be strongly occluded by other voxels closer to the camera. This posterior expression expands to,

$$P^{N+!}\left(X \in S\right) = P^N\left(X \in S\right)\frac{\sum\limits_{X_\alpha \in R_{x^{N+1}}} p^N\left(I_x^{N+1} \mid V_{x^{N+1}} = X_\alpha\right) P^N(V_{x^{N+1}} = X_\alpha \mid X \in S)}{\sum\limits_{X_\alpha \in R_{x^{N+1}}} p^N(I_x^{N+1} \mid V_{x^{N+1}} = X_\alpha)\, P^N(V_{x^{N+1}} = X_\alpha)}\,.$$

The effect of occlusion comes into play in the term, $P^N(V_{x^{N+1}} = X_\alpha)$. This expression represents the probability that the observed image intensity could arise from voxel, X_α. This probability in turn is composed of two parts: 1) voxel X_α contains a surface element; and voxel X_α is not occluded. The probability that the voxel is not occluded (i.e. visible) is given by,

$$\prod_{\alpha' < \alpha}\left[1 - \mathrm{P}(X_{\alpha'} \in S)\right],$$

and so,

$$P^N(V_{x^{N+1}} = X_\alpha) = \mathrm{P}(X_\alpha \in S)\prod_{\alpha' < \alpha}\left[1 - \mathrm{P}(X_{\alpha'} \in S)\right].$$

The mixing weight increment, $d\omega_k^{N+1}$ is assigned to voxels along the ray in proportion to the relative values of $P^N(V_{x^{N+1}} = X_\alpha)$.

As this inference process is repeated over a sequence of images, both the surface probabilities and appearance mixtures will converge to values that optimally account for the observations. It has been shown that the process is guaranteed to converge provided that there is at least one voxel in the scene that is unoccluded in a sufficient number of views[2]. This condition is easily met in aerial images. An example of the surface probabilities in a volume after training on a short video sequence is shown in Figure 4. The figure illustrates the probabilistic nature of the scene geometry, which reflects the evidence available from the image sequence. It is interesting to note that vegetation is relatively well captured, even though its appearance in different images is variable due to differing internal shadowing and occlusion. The light plume in the background is due to observation of image pixels that correspond to 3-d points that lie outside the volume of interest.

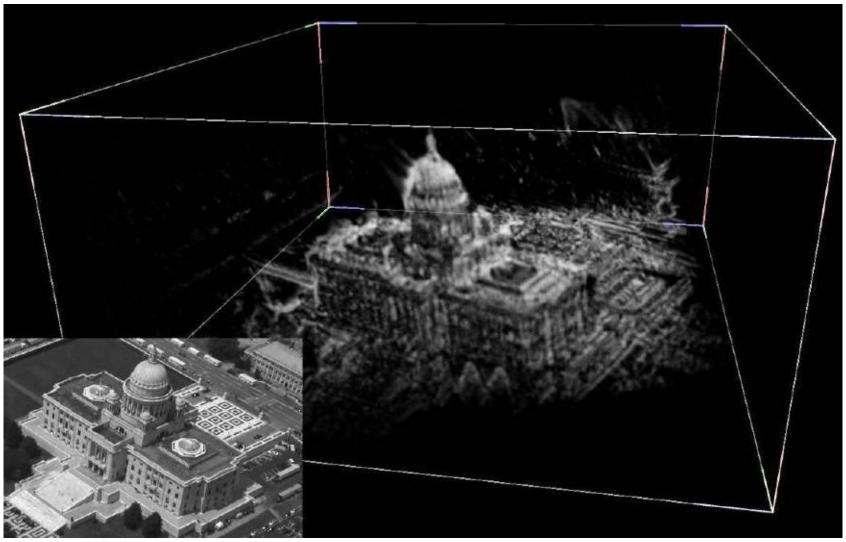


Figure 4 A volume rendering of surface probabilities learned from a short aerial video sequence. A typical video frame is shown in the inset. In the rendering, voxels with a high surface probability are more opaque than voxels that do not contain surface elements. (Image from the thesis by T. Pollard, Brown University[3].)

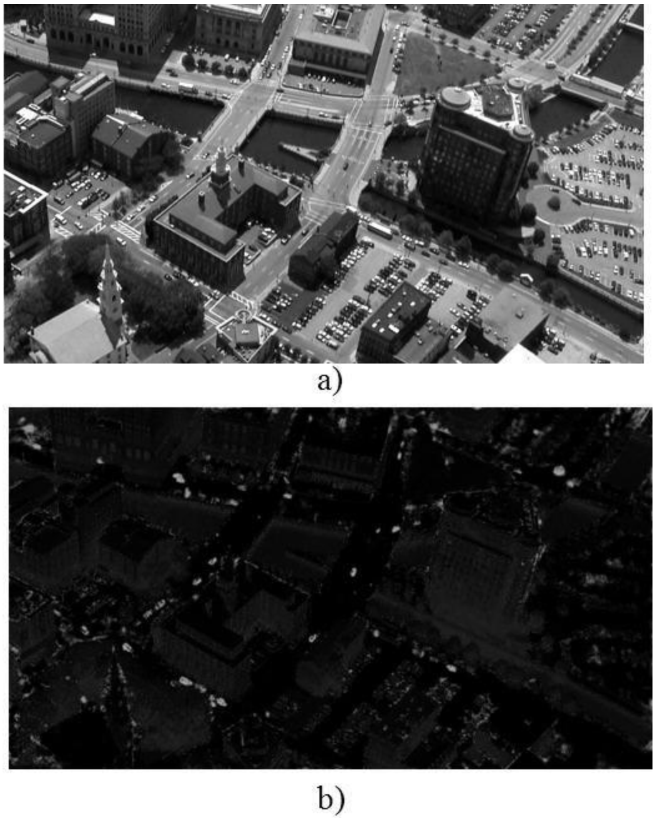


Figure 5 The detection of changes in an aerial video. a) A typical video frame. b) The probability of change, or foreground (shown in white). The changes are due to moving vehicles. Note that false changes due to occlusion of 3-d building structure are negligible.

2.3 Background modeling and change detection

The probabilistic framework just described is capable of detecting changes in a new image with respect to variations that have normally occurred in a prior sequence of images. The process of change detection is carried out by determining the probability of each pixel in a new image, given the volumetric appearance model that has been updated by a series of prior images. The probability density for the intensity of a pixel at location x is given by,

$$p(I_x \mid x \in b) = \sum_{X_\alpha \in R_x} p(I_x \mid V_x = X_\alpha)\, P\ (V_x = X_\alpha)\ ,$$

where R_x is the ray through pixel location x. The predicate, $x \in b$, indicates that the probability density is conditioned on the current state of the volumetric model, which represents normal background. This probability density can be turned into a probability by the following Bayesian posterior calculation.

$$P(x \in b \mid I_x) = \frac{p(I_x \mid x \in b) P(x \in b)}{p(I_x \mid x \in b) P(x \in b) + p(I_x \mid x \in f)\left[1 - P(x \in b)\right]}$$

The prior probability that a given image pixel location is background is, $P(x \in b)$. If a new object has appeared at location x that was not present in the images used to train the voxel model then that object is considered to be *change*, also called *foreground*. The posterior estimate of change probability requires the probability density for observed intensities (or colors) of foreground objects. A reasonable choice for this density is a uniform distribution over the full range of possible intensities, e.g. (0, 255). An example of $P(x \in b \mid I_x)$ is shown in Figure 5, where b) illustrates $P(x \in f \mid I_x) = 1 - P(x \in b \mid I_x)$. The model in this example was trained on 90 video frames of an urban scene. Note that most of the vehicles have been clearly segmented from the background since they are abnormal with respect to the appearance of the street surfaces. At any given point on the roadway the majority of training samples for the volumetric model are street surface colors. Thus, given the volumetric appearance model it is possible to segment moving objects in a video or in a series of still overhead images, such as satellite data.

3. CLASSIFYING FOREGROUND OBJECTS: VEHICLES

Based on the previous analysis it can be assumed that estimates of foreground probabilities are available from the voxel appearance model. Objects of interest can be detected as regions of high foreground probability. Such foreground objects can arise under a wide range of circumstances such as new construction, weapon damage, vegetation growth and vehicle motion. This paper will focus on the latter category due to its high level of importance for many surveillance and intelligence applications. It can be assumed that pixels on vehicles that are present in a new image will have significantly higher values of $P(x \in f \mid I_x)$ than stationary structures, except for other types of genuine change.

3.1 Foreground vehicle appearance operators

The focus of this investigation is on vehicle classification that is observed at relatively low spatial resolution, e.g. 25-100 pixels on the vehicle. At this resolution, it has been determined that an effective feature for representing vehicle appearance is a family of oriented intensity extrema operators described by the following kernel,

$$k(u,v) = \frac{1}{2\pi\lambda_0\lambda_1^3}\left(\frac{(uSin\theta + vCos\theta)^2}{\lambda_1^2} - 1\right) e^{-\frac{1}{2}\left(\left(\frac{Cos^2\theta}{\lambda_0^2} + \frac{Sin^2\theta}{\lambda_1^2}\right)u^2 + \left(\frac{Sin^2\theta}{\lambda_0^2} + \frac{Cos^2\theta}{\lambda_1^2}\right)v^2 + 2uv\, Sin\theta\, Cos\theta\left(\frac{1}{\lambda_0^2} - \frac{1}{\lambda_1^2}\right)\right)},$$

where λ_0 and λ_1 define the scale of the operator and θ its orientation. The vector of these parameters is defined by the symbol $\alpha = \{\lambda_0, \lambda_1, \theta, \beta\}$, where β specifies that the operator responds to either dark or bright extrema.

An example of applying this operator to vehicle foreground is shown in Figure 6. In this example the operators are detecting the contrast between the road surface and the vehicle, as well as any shadow cast by the vehicle. In practice, series of such operators is applied at different orientations with both bright and dark extrema detection signs.

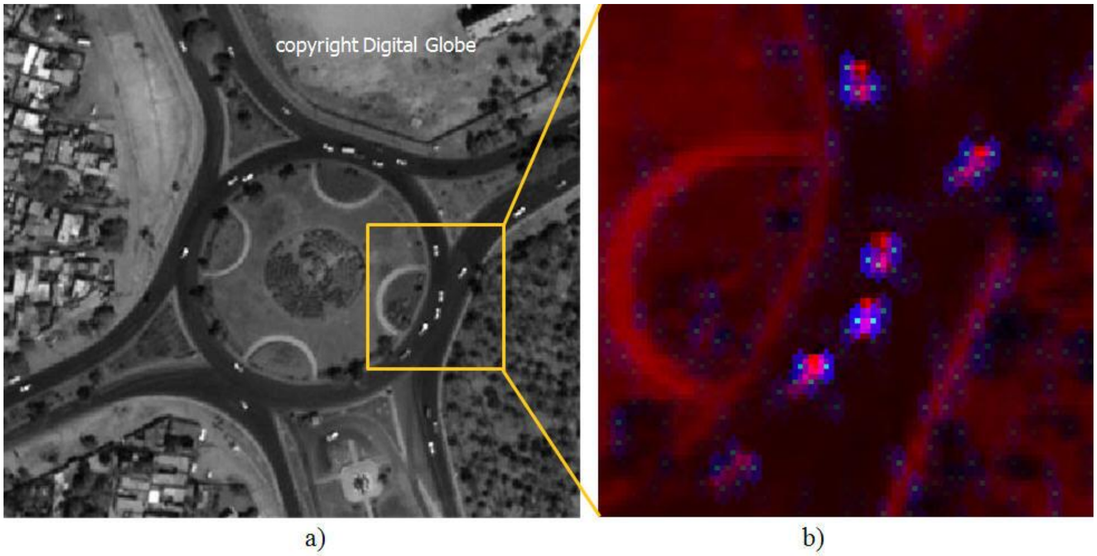


Figure 6 An example of vehicle extrema operator responses. $\alpha = \{1, 0.5, 90^o, dark\}$. The spatial resolution is around 0.7 meters, with about 25 pixels on a vehicle. The operator response is indicated by the cyan dot. The operator kernel extent is indicated in blue. The original grey scale intensity is in the red channel.

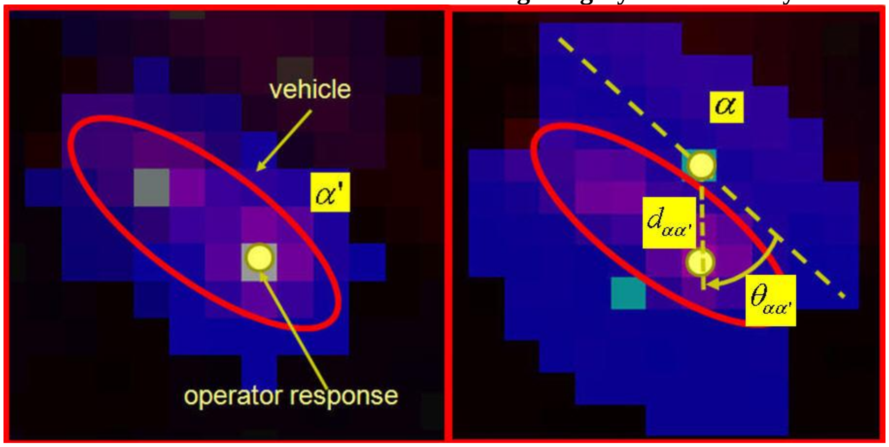


Figure 7 The composition of extrema operators. The anisotropic dark operator, α , is composed with one of a bright peak operator, α' . The composition is characterized by distance $d_{\alpha\alpha'}$ and relative orientation $\theta_{\alpha\alpha'}$.

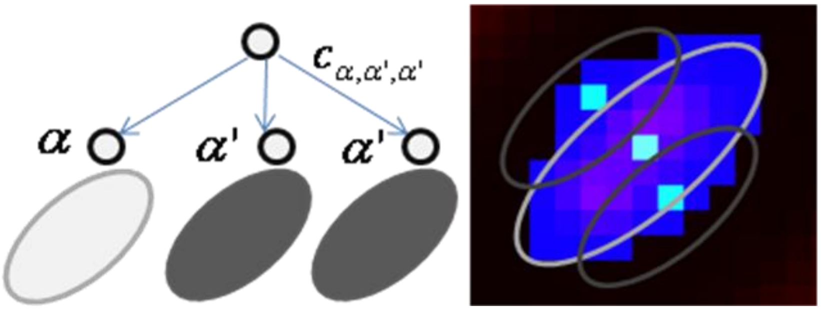


Figure 8 Three primitive extrema operators compose in a Layer 1 node. The central part is $\alpha = \{2, 1, -45^o, bright\}$, and the second primitive part is $\alpha' = \{2, 1, -45^o, dark\}$. The peak responses of the operators are indicated by cyan pixels. The operator kernel is indicated in blue. The vehicle intensity is in the red channel.

3.2 A compositional model for vehicle classification

There is growing interest in models for object classification based on hierarchical composition[4,5]. In this approach, a hierarchy of *parts* is defined, starting with primitives such as the extrema operators defined in the previous section. Each compositional layer combines parts from previous layers. An example of a first level composition of the extrema operators is shown in Figure 7. In this example, two operators, α and α' are geometrically related by the distance between the operator response peak locations, $d_{\alpha\alpha'}$, and the relative orientation of the vector joining the response peaks, $\theta_{\alpha\alpha'}$. This paired combination forms a new component (part) that can be combined with other parts to build up a complete description of the vehicle category. Let the composition of operators α and α' be denoted as $c_{\alpha\alpha'}$. The posterior probability of the existence of such a composition is given by Bayes law,

$$P(c_{\alpha\alpha'} \mid d_{\alpha\alpha'}, \theta_{\alpha\alpha'}) = \frac{p(d_{\alpha\alpha'}, \theta_{\alpha\alpha'} \mid c_{\alpha\alpha'}) P(c_{\alpha\alpha'}) P(\alpha) P(\alpha')}{p(d_{\alpha\alpha'}, \theta_{\alpha\alpha'})}$$

$P(\alpha)$ and $P(\alpha')$ are the probabilities that the primitive operator responses have been observed and $P(c_{\alpha\alpha'})$ is the prior probability of composition $c_{\alpha\alpha'}$. The density, $p(d_{\alpha\alpha'}, \theta_{\alpha\alpha'} \mid c_{\alpha\alpha'})$ is acquired from manually selected samples of occurrences of the operator composition on labeled vehicle instances. The denominator is the marginal density taken over all compositions of α and α', including the case where no coherent composition exists. The probabilities $P(\alpha)$ and $P(\alpha')$ are acquired through Bayesian learning, where the process takes into account the existence of foreground and background probabilities described earlier. This automatic learning process is now described.

3.3 Learning operator probabilities

The probability of the occurrence of a given operator type, $P(\alpha)$, depends on whether it occurs on a foreground vehicle object or background. The analysis can be formulated in terms of the strength of the operator response on both foreground and background. Denote the response of an operator of type α as o_α occurring at pixel location, x_α. There are four possible hypotheses concerning the interpretation of location x_α: $x_\alpha \in f \wedge x_\alpha \in v$, x_α is a foreground pixel and also on a vehicle; $x_\alpha \in f \wedge x_\alpha \notin v$, x_α is a foreground pixel and not on a vehicle; $x_\alpha \in b \wedge x_\alpha \in v$, x_α is a background pixel and on a vehicle; and $x_\alpha \in b \wedge x_\alpha \notin v$, x_α is a background pixel and not on a vehicle. As an example, the posterior probability of the operator, α, existing as foreground at a vehicle, is given by,

$$P(x_\alpha \in f \wedge x_\alpha \in v \mid o_\alpha) = \frac{p(o_\alpha \mid x_\alpha \in f \wedge x_\alpha \in v)}{p(o_\alpha)} P(x_\alpha \in f) P(x_\alpha \in v),$$

assuming that the existence of a vehicle at x_α is independent of the location being foreground. In order to compute the posterior probabilities for the four hypotheses it is necessary to model and estimate the four operator response densities:

a) $p(o_\alpha \mid x_\alpha \in f \wedge x_\alpha \in v)$
b) $p(o_\alpha \mid x_\alpha \in f \wedge x_\alpha \notin v)$
c) $p(o_\alpha \mid x_\alpha \in b \wedge x_\alpha \in v)$
d) $p(o_\alpha \mid x_\alpha \in b \wedge x_\alpha \notin v)$

The first two densities can be learned from a training set of labeled vehicle instances, or if it is assumed that the majority of foreground objects are vehicles, then distribution a) can be learned without labeling, i.e. unsupervised. Since vehicles are moving objects it is reasonable to consider distribution a) to be independent of location x_α. The foreground density is modeled as a Weibull distribution,

$$p(o_\alpha \mid x_\alpha \in f \wedge x_\alpha \in v) = \frac{k}{\lambda}\left(\frac{o_\alpha}{\lambda}\right)^{k-1} e^{-\left(\frac{o_\alpha}{\lambda}\right)^{k_f}} ,$$

where k is the shape parameter and λ the scale parameter. The Weibull parameters are acquired by fitting the distribution to the histogram of foreground operator responses. The distribution b) is also considered independent of location x_α and modeled as a Weibull distribution. Its parameters are acquired by fitting the distribution to the histogram of operator responses that fire on the non-vehicle regions of the labeled images for the supervised case, and on the no-change regions for the unsupervised case. This case represents appearance of the operators for non-class regions/objects.

The background cases, c) and d), are merged into the single distribution, $p(o_\alpha \mid x_\alpha \in b)$, which is learned for each location in the scene. The background distributions are learned from the expected normal background image constructed from the voxel appearance models as described in Section 2.3. The expected intensity at location x is given by,

$$\left\langle I_x \right\rangle = \frac{\int I_x p(I_x \mid x \in b)\, dI}{\int p(I_x \mid x \in b)\, dI} .$$

It is also possible to estimate the standard deviation in this expected intensity, denoted by σ_x, by sampling the Gaussian mixtures along the camera ray through location x. The background operator response distribution is modeled by a Gaussian,

$$p(o_\alpha \mid x_\alpha \in b) = \frac{1}{\sqrt{2\pi}\sigma_x^\alpha} e^{-\frac{1}{2}\left(\frac{o_\alpha - \mu_x^\alpha}{\sigma_x^o}\right)^2} .$$

The values μ_x^α are computed by applying the operator kernel to the expected image, $\left\langle I_x \right\rangle$. The values of σ_x^α are computed according to,

$$\left(\sigma_x^\alpha\right)^2 = \sum_{k=-r_k}^{k=+r_k} \sum_{l=-r_l}^{l=+r_l} w_{k,l}^2 \, \sigma_x^2 (i+l, j+k),$$

where image location $x = (i, j)$ and $w_{k,l}$ are the operator kernel weights. The motivation for using a Gaussian distribution for background is that the classification into bright and dark operator responses can be determined with some finite probability even though the response on the expected image is of the opposite sense. That is, the Gaussian tail extends across the zero operator response boundary. The current vehicle classifier does not attempt to identify vehicles in the background, e.g., vehicles in a parking lot are ideally ignored, provided their appearance is well-represented in the expected background image, $\left\langle I_x \right\rangle$.

In summary, the 3-d voxel model supports automated learning of operator distributions for the normal background scene as well as moving vehicles that are automatically attributed with high foreground probabilities.

4. EXPERIMENTAL RESULTS

4.1 Quick Bird satellite imagery

The first evaluation of the modeling and recognition system is based on a set of Quick Bird satellite images with 0.7 meters pixel resolution, about 25 pixels on a vehicle. The voxel world is constructed using 11 images of a region of interest (ROI), and tested on a novel view of the region, Figure 10 a). An expected image is generated from voxel world for this view, Figure 10 b), and is used for background response density estimation. The detector hierarchy shown in Figure 8 was manually designed to account for typical arrangements of extrema operators. The performance of both supervised and unsupervised training for foreground response model fitting is examined and given in Figure 9 a). Figure 10 d) shows the output map of the detector which is the posterior for a composition to be a vehicle foreground:

$$P(c^{v,f}_{\alpha\alpha} \mid d_{\alpha\alpha'}, \theta_{\alpha\alpha'}, o_\alpha, o_{\alpha'}) = \frac{p(d_{\alpha\alpha'}, \theta_{\alpha\alpha'} \mid c^{v,f}_{\alpha\alpha'}) P(c^{v,f}_{\alpha\alpha'}) p(o_\alpha \mid x_\alpha \in f \wedge x_\alpha \in v) p(o_{\alpha'} \mid x_{\alpha'} \in f \wedge x_{\alpha'} \in v)}{p(d_{\alpha\alpha'}, \theta_{\alpha\alpha'}, o_\alpha, o_{\alpha'})}$$

Figure 10 e), f) and g) shows the posteriors for the remaining three cases: non-vehicle foreground, $c^{\bar{v},f}_{\alpha\alpha'}$, vehicle background, $c^{v,b}_{\alpha\alpha'}$, and non-vehicle background, $c^{\bar{v},b}_{\alpha\alpha'}$. The denominator is the sum of all four cases.

For vehicle compositions, the variations of distance and the angle are modeled as independent Gaussian distributions. These distributions are fit to a set of labeled exemplars of the class during training, so that

$$p(d_{\alpha\alpha'}, \theta_{\alpha\alpha'} \mid c^{v,f}_{\alpha\alpha'}) = N(d_{\alpha\alpha'}; c_{\alpha\alpha'}) N(\theta_{\alpha\alpha'}; c_{\alpha\alpha'}) = p(d_{\alpha\alpha'}, \theta_{\alpha\alpha'} \mid c^{v,b}_{\alpha\alpha'}) .$$

For non-vehicle compositions, the operators are assumed to fire random in the receptive field of the composition, thus the variations of distance and angle are modeled as uniform distributions, so that

$$p(d_{\alpha\alpha'}, \theta_{\alpha\alpha'} \mid c^{\bar{v},f}_{\alpha\alpha'}) = U(d_{\alpha\alpha'}; r) U(\theta_{\alpha\alpha'}; r) = p(d_{\alpha\alpha'}, \theta_{\alpha\alpha'} \mid c^{\bar{v},b}_{\alpha\alpha'}) ,$$

where r denotes the radius of the receptive field and it doubles with each layer in the hierarchy. There are four parameters of the system regarding the prior probabilities used in the equations; they are set as follows in these experiments:

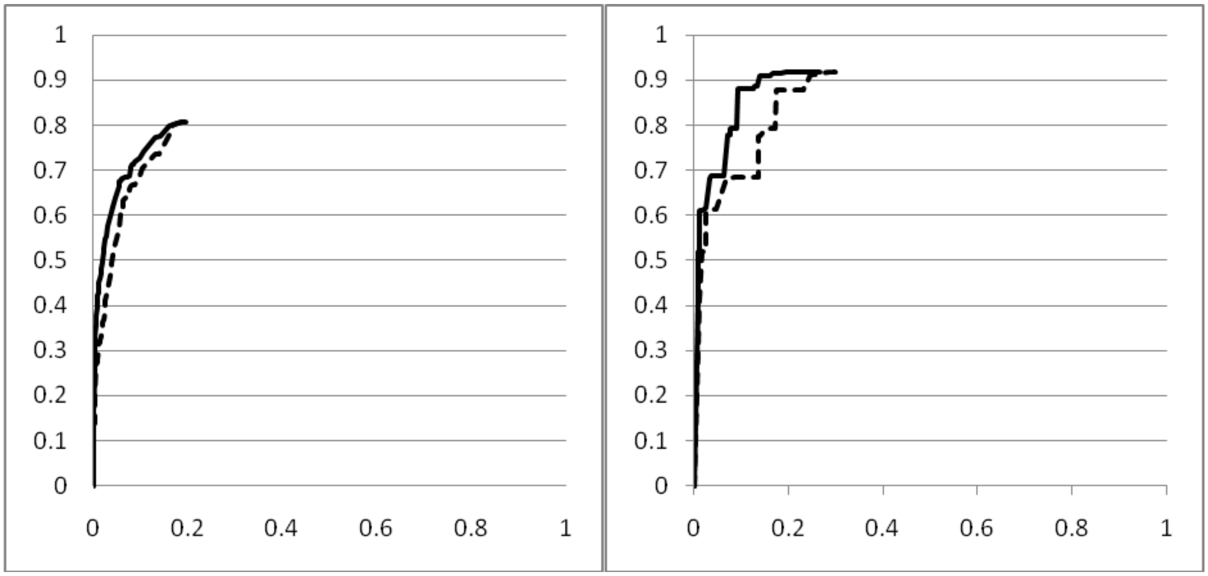
$$P(x_\alpha \in v) = 0.1, P(c^{v,f}_{\alpha\alpha'}) = 0.15, P(c^{\bar{v},f}_{\alpha\alpha'}) = 0.15, P(c^{v,b}_{\alpha\alpha'}) = 0.15 .$$

Note from Figure 9 a) that the reduction in recognition performance due to unsupervised learning is very small. This result indicates that most of the foreground objects are vehicles, and so manual labeling of ground truth is not needed. Note from Figures 8 e) and g) that there are many primitives having non-vehicle appearance that compose randomly and these are successfully separated from the true vehicle-foreground detections shown in Figure 10 d). This demonstrates the contribution of having vehicle-non-vehicle distinction in our appearance and geometry models in reducing the false alarm rate.

4.2 Aerial video

The beneficial effect of voxel background modeling is evaluated on an aerial video sequence where there are parked vehicles as well as moving vehicles in the region of interest. This video imagery has resolution of about 100 pixels on a vehicle. Two compositions of extrema operators are designed to detect both bright and dark vehicles using predominantly bright and dark primitives respectively. Figure 9 b) shows the ROC for the classifier based on the posterior probability for both compositions. The solid curve is the performance for vehicle-foreground case, making use of foreground probabilities estimated from the voxel model, shown in Figure 11 d). The dashed curve is the performance for the vehicle composition posterior only, shown in Figure 11 e). In this case, the foreground-background distinction is eliminated by setting the foreground probability to one at each position. Note that there is a significant decrease in the false alarm rate when the foreground probability map is introduced. However, it can be observed in Figure 11 c) that the background probability over parked vehicles is lower than over open regions with low texture. This effect is partially due to insufficient spatial resolution of the voxel grid, so that vehicles are not well resolved in the expected image.

Development is under way to implement a multi-resolution grid based on an octree index. This approach should produce a significantly better quality expected image for high-resolution video.



a) b)

Figure 9 a) ROC for the test view of Quick Bird satellite imagery. The solid curve shows recognition performance when the foreground response models of the primitive parts are trained with the ground-truth map of one of the training views (supervised case) and dashed curve shows performance for training using locations of high foreground probability (unsupervised case). b) ROC for the test view of the aerial video. The solid curve indicates recognition performance using foreground probabilities from the voxel model. The dashed curve indicates recognition performance when the background model is not used. The voxel model eliminates a significant number of false alarms on parked vehicles.

REFERENCES

[1] Stauffer, C. and Grimson, W.E.L., “Adaptive Background Mixture Models for Real-Time Tracking,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Vol 2, 245-252, 1999.

[2] Pollard, T. and Mundy, J. L., “Change Detection in a 3-d World,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 1-6, 2007.

[3] Pollard, T., “Comprehensive 3-d change detection using volumetric appearance modeling,” Ph.D. Thesis, Brown University, 2008.

[4] Jin, Y. and Geman, S., “Context and Hierarchy in a Probabilistic Image Model,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Vol 2, 2145-2152, 2006.

[5] Fidler, S. and Leonardis, A., “Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 1-8, 2007.

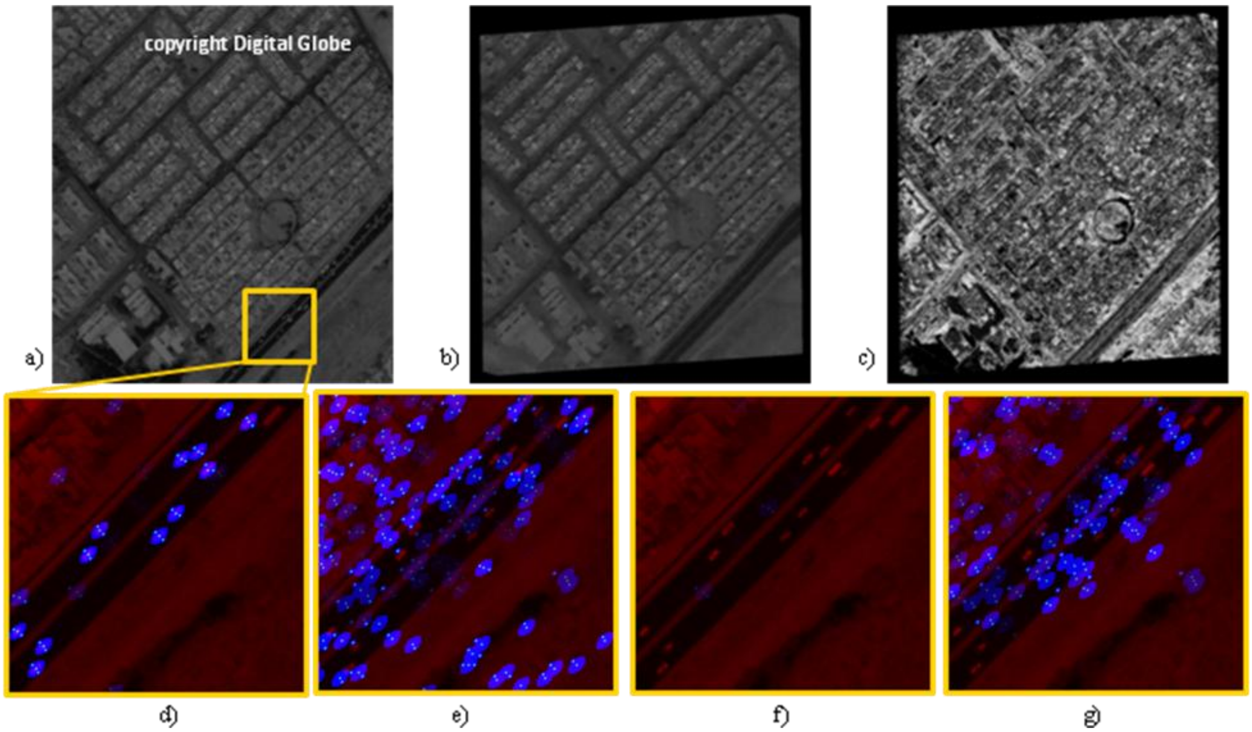


Figure 10 a) The test view of ROI. b) The expected image generated by the voxel world for the test view. c) The estimated background probability map for the test view. d), e), f) and g) are the posterior composition probabilities for the vehicle-foreground, non-vehicle foreground, vehicle-background and the non-vehicle background respectively.

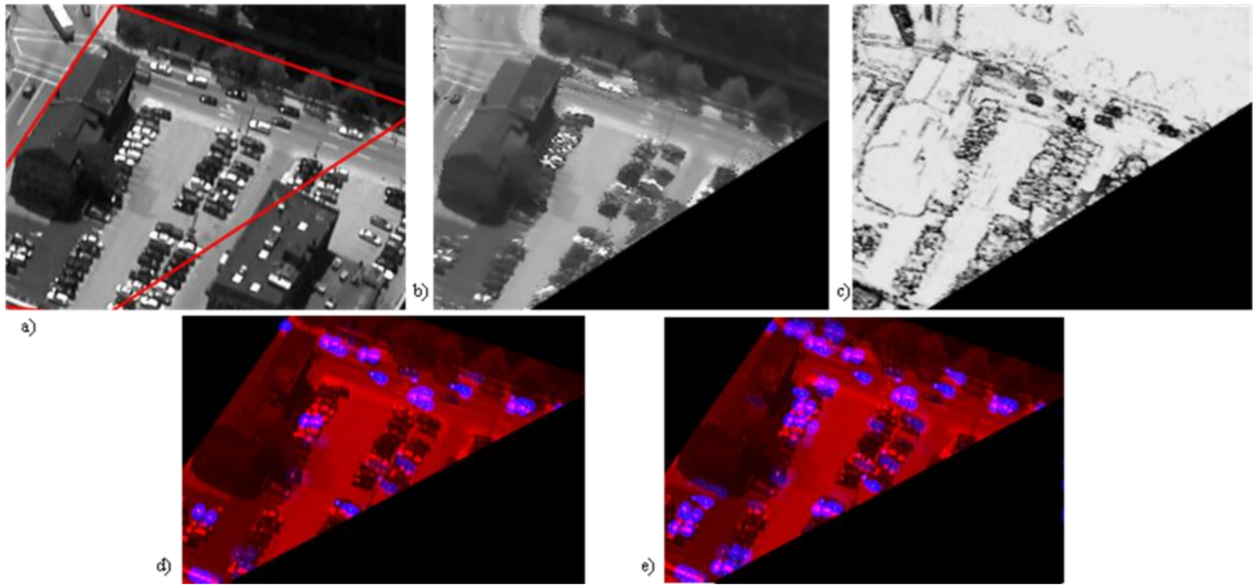


Figure 11 a) ROI marked on a typical video frame. b) The expected image generated by the voxel world. Note that only parked vehicles appear, since moving objects are averaged out. c) The background probability map of the image in a). d) The posterior vehicle probability based on the background voxel model (blue) e) the posterior when the background model is not used. Note the significant increase in false alarms (parked vehicles) in the parking area.