Analysis of Persistent Motion Patterns Using the 3D Structure Tensor

Robert Pless
Department of Computer Science
Washington University in St. Louis
pless@cs.wustl.edu

John Wright
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
jnwright@uiuc.edu

Abstract

Surveillance applications often capture video over long time periods; interpretation of this data is facilitated by background models that effectively represent the typical behavior in the scene. Capturing statistics of the spatio-temporal derivatives at each pixel can efficiently model surprisingly complicated motion patterns. Considering the video as a function of space and time, the mean 3D structure tensor at each pixel characterizes local image variation, the most common local motion, and whether that motion is consistent or ambiguous. Furthermore, this structure tensor field — the structure tensor at each pixel — is interpretable as a constrained Gaussian probability density function over the derivatives measured across the entire image. In scenes with multiple global motion patterns, a mixture model (of these global distributions) automatically factors background motion into a set of flow fields corresponding to the different motions. The models are developed online in real time and can adapt to changes in background motion. We demonstrate the ability to automatically discover the different motion patterns in an intersection.

1 Introduction

Online detection of motion patterns in video footage is useful in a variety of applications, such as visual surveillance, dynamic background subtraction and real-time video compression. In this paper, we address the question of detecting and modeling persistent patterns of background motion. A model of common motion patterns allows higher-level processes to ignore common motions in the scene, and provides a natural basis for discovering anomalous objects, the essential first step in any automatic surveillance system.

The problem of statistical background modeling has been approached mostly in the context of visual surveillance and anomaly detection. In one of the first papers to move beyond naive frame-differencing or constant background approaches, Stauffer and Grimson use a Gaussian mixture model to approximate a multi-modal color distribution at each pixel. This method is highly successful on static and quasi-static backgrounds [10]. While their model does not attempt to handle backgrounds that contain regular motion, it does demonstrate the power of mixture models as a real-time background modeling technique.

Another notable system, W4, uses a background model consisting of a maximum and minimum intensity value at each pixel, along with the maximum temporal derivative [3]. W4's background model produces a crude initial segmentation that is then postprocessed by code specific to finding and recognizing humans. While it is true that background modeling is usually just a first step in a larger system, we believe that local spatio-temporal statistics are powerful enough to provide both intelligent anomaly detection and more useful semantic cues for higher level processes.

Recently, there has been increased interest in background subtraction in dynamic scenes. It is widely recognized that background motions such as waving trees and ocean waves create insurmountable difficulties for conventional algorithms. Moreover, while the definition of "background" varies from application to application, there are certainly times when it is desirable to consider cars obeying traffic laws or pedestrians following common patterns part of the background.

Surprisingly, most of the attempts to model dynamic backgrounds have been appearance-based, rather than motion-based. Zhong and Sclaroff treat the entire video as a single dynamic texture, classifying anomalous pixels as those in which the current frame deviates from the prediction of a Kalman filter defined on the coefficients of a PCA reconstruction [11]. Monnet, et al. perform incremental PCA on image blocks and then classify anomalies based on the magnitude of the component normal to the first N eigenvectors [7]. Because they are solely intensity-based, both of these methods fail in video sequences where the motion is repetitive, but image appearance is not. For example, the motion of a car in traffic is a strong indicator of whether it is behaving anomalously, while its color is not. Futhermore, in most traffic sequences, there is no discernable pattern to the color of the cars, rendering time-series prediction of the appearance of a given frame difficult at best.

A few prior works have discussed methods that model motions in background scenes. Mittal and Paragios model

background statistics by combining optic flow (computed from image derivatives) and color cues in a kernel density estimation framework [6]. While this method does explicitly model image motion, Pless et al. have success modeling background motions simply as distributions of spatio-temporal image derivatives at each pixel [9].

This paper describes the first real-time method for factoring background motions into multiple flow fields, based on the aggregation of simple local statistics in the form of one or more tensor fields. Furthermore, we introduce the use of the 3-D structure tensor in background modeling as a convenient tool for representing the joint distribution of x, y, and t derivatives at each pixel. The structure tensor has been applied to a variety of problems local video analysis such as road detection [8], specularity removal [2], and motion estimation [5]. This work illustrates the use of the structure tensor field as tool for representing global patterns of local motions, which may be of broader interest.

Next, we review useful properties of the structure tensor, including its relationship to the Gaussian distribution of spatio-temporal derivatives. Section 3 discusses the structure tensor field as a global model and methods for factoring streaming video data that arises from different global motion patterns. Finally, Section 4 show results from real-time analysis of an intersection, showing surprisingly accurate decomposition into flow fields.

2 The Structure Tensor Field

The structure tensor has been widely used in the image analysis field for optic flow estimation and segmentation. In these cases, the structure tensor at a pixel is defined based on the spatio-temporal image derivatives measured in a region around that pixel [1]. In the case of surveillance, the camera is stationary and viewing a scene containing motion patterns that may be consistent or recur over long time periods, it is reasonable to combine data through time instead of over a region in the image.

2.1 The Structure Tensor

Let $\nabla I(\vec{p},t) = (I_x(\vec{p},t), I_y(\vec{p},t), I_t(\vec{p},t))^T$ be the spatio-temporal derivatives of the image intensity $I(\vec{p},t)$ at pixel $\vec{p}$ and time t. At each pixel, the structure tensor, Σ, is defined as

$$\Sigma(\vec{p}) = \frac{1}{f}\sum_{t=1}^{f} \nabla I(\vec{p},t)\nabla I(\vec{p},t)^T$$

where f is the number of frames in the sequence and $\vec{p}$ is omitted after this for clarity's sake. Except as described in section 3.2, we consider these distributions to be independent at each pixel. To focus on scene motion, the measurements are filtered, only considering measurements that come from motion in the scene, that is, measurements for which $|I_t| > 0$. For the sake of the clarity of the exposition in section 2.2, we assume the mean of ∇I to be zero (which does *not* imply the motion is 0).

Under this assumption, Σ defines a Gaussian distribution $\mathcal{N}(0,\Sigma)$. Previous work in anomaly detection [9] can be cast nicely within this framework: anomalous measurements can be detected by comparing either the mahalanobis distance, $\nabla I^T \Sigma^{-1} \nabla I$, or the negative log-likelihood $ln((2\pi)^{3/2}|\Sigma|^{1/2}) + \frac{1}{2}\nabla I^T \Sigma^{-1} \nabla I$, to a preselected threshold [9].

In real-time applications, computing with the entire sequence is not feasible and the structure tensor must be estimated online. Assuming the distribution is stationary, Σ can be estimated as the sample mean of $\nabla I \nabla I^T$,

$$\Sigma_t = \frac{(n-1)}{n}\Sigma_{t-1} + \frac{1}{n}\nabla I \nabla I^T$$

However, it is unrealistic to assume that the distribution at a given pixel will be stationary throughout an entire video sequence. The model can be allowed to drift by instead assigning a constant weight, $\alpha \in [0,1]$, to each new measurement:

$$\Sigma_t = (1-\alpha)\Sigma_{t-1} + \alpha \nabla I \nabla I^T.$$

This update method causes the influence of a given measurement on Σ to decay exponentially, with decay constant $\frac{-1}{ln(1-\alpha)}$.

2.2 Relationship to 2-D Image Motion

The structure tensor field's value as a background model comes from the strong relationship between optic flow and the spatio-temporal derivatives, via the optic flow constraint equation, $I_x u + I_y v + I_t = 0$ [4]. This equation constrains all gradient measurements produced by a flow (u,v) to lie on a plane through the origin in I_x, I_y, I_t-space. The optic flow vector, $(u,v,1)$, is normal to this plane.

Suppose the distribution of ∇I measurements comes from different textures with the same flow, and one models this distribution as a Gaussian, $\mathcal{N}(0,\Sigma)$. Let $\vec{x_1}, \vec{x_2}, \vec{x_3}$ be the eigenvectors of Σ and $\lambda_1, \lambda_2, \lambda_3$ the corresponding eigenvalues. Then $\vec{x_1}$ and $\vec{x_2}$ will lie in the optic flow plane, with $\vec{x_3}$ normal to the plane and $\lambda_1, \lambda_2 \gg \lambda_3$. In fact, it can be shown that the $\vec{x_3}$ is the total least-squares estimate of the homogeneous optic flow, $\frac{(u,v,1)}{\|(u,v,1)\|}$.

The covariance matrix permits deeper analysis than just computing an estimate of the optic flow at each pixel. One simple confidence measure for the optic flow estimate is

$\mathbf{S} = 1 - \lambda_3/\lambda_2$. This measure takes on values in $[0, 1]$ and is large when the second eigenvalue is much larger than the third. When the texture along the direction of motion is insufficient to uniquely determine the true optic flow (the aperture problem), the Gaussian distribution will appear long and thin, with comparable λ_2 and λ_3. Since the distribution does not provide a strong indication of the true orientation of the optic flow plane, our confidence in the total least squares solution, $\vec{x_3}$, should be low. $\mathbf{S}$ satisfies this intuition, since it approaches 0 as λ_3 approaches λ_2.

Figures 1 and 2 show estimated flow fields for several scenes, along with plots of $\mathbf{S}$. Note that in areas with multiple motions, $\mathbf{S}$ drops dramatically. This is indicative of the fact that the distribution of derivatives from two different motions does not fall on a plane in the spatio-temporal derivative space. The next section describes how the model can be expanded to handle scenes with multiple motions at a given pixel — by generating multiple global motion patterns.

3 Multiple Structure Tensor Fields

3.1 A Single Joint Distribution

The previous section showed how a structure tensor field defines a zero-mean Gaussian at each pixel. If the inter-pixel covariances are constrained to be zero, this set of distributions may be considered as a single joint Gaussian, $\mathcal{N}_{global}$ over the entire image. Let Σ_i be the structure tensor at the i-th pixel. Then the covariance matrix of the global distribution is the block-diagonal matrix

$$\tilde{\Sigma}_{global} = \begin{pmatrix} \Sigma_1 & & \mathbf{0} & \\ & \Sigma_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \Sigma_p \end{pmatrix}$$

Let $\tilde{\nabla} I$ be the concatenation of the gradient vector at each individual pixel: $\tilde{\nabla} I = (I_x^{(1)}, I_y^{(1)}, I_t^{(1)}, I_x^{(2)}, \ldots)$. Then the likelihood of the observation at a given frame is

$$P(\tilde{\nabla} I | \mathcal{N}_{global}) = k \exp(-\frac{1}{2} \tilde{\nabla} I^T \tilde{\Sigma}_{global}^{-1} \tilde{\nabla} I)$$

where k is a normalizing constant. Because $\tilde{\Sigma}$ is block diagonal, this can be rewritten as:

$$P(\tilde{\nabla} I | \mathcal{N}_{global}) = \prod_i P(\nabla I_i | \mathcal{N}_i(0, \Sigma_i)).$$

3.2 Mixture Models of Joint Distributions

In Section 2.2, we discussed the strong relationship between the structure tensor and the optic flow at each pixel. This leads directly to a relationship between the optic flow field and the constrained Gaussian distribution over all derivative measurements.

The background model can therefore be modified to handle multiple motions. Each motion field is treated as a joint Gaussian distribution over the entire image as described above. These large Gaussians are combined in a single mixture model,

$$w_1 \mathcal{N}_1(0, \tilde{\Sigma}_1) + \ldots + w_M \mathcal{N}_M(0, \tilde{\Sigma}_M) + w_{unk} \mathcal{M}_{unk}$$

where M is the number of unique background motions. $\mathcal{M}_{unk}$ is the prior distribution of (I_x, I_y, I_t) vectors for motions not fitting any background model – including anomalous events and objects that do not follow the background. $\mathcal{M}_{unk}$ may be chosen as a uniform distribution, or as an isotropic Gaussian, with little qualitative effect on the mixture estimated. One advantage of choosing a uniform foreground prior is that anomalous objects can be detected by simply thresholding the negative log-likelihood of the backgrounds.

The model is a Gaussian mixture model and can be updated according to the standard adaptive mixture model update equations (as used, for example, in [10]), although here it is applied to a very high-dimensional distribution. The special block-diagonal structure simplifies the computations. The mixture model can be updated online, by first calculating the likelihoods:

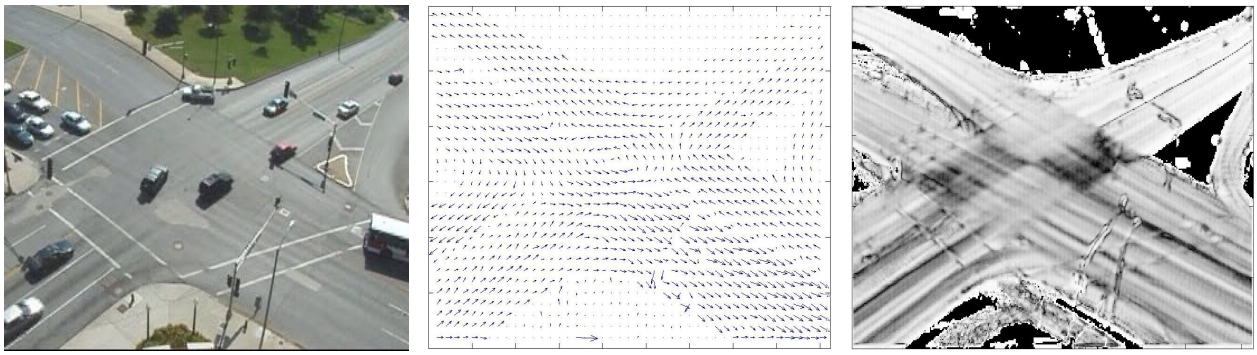
$$P(\mathcal{N}_i | \tilde{\nabla} I) = \frac{w_i P(\tilde{\nabla} I | \mathcal{N}_i)}{w_{unk} P(\tilde{\nabla} I | \mathcal{M}_{unk}) + \sum_{j=1}^{M} w_j P(\tilde{\nabla} I | \mathcal{N}_j)}$$

Each of the fields can then be updated as:

$$\tilde{\Sigma}_{i,t} = (1 - \beta_i) \tilde{\Sigma}_{i,t-1} + \beta_i \tilde{\nabla} I \tilde{\nabla} I^T$$

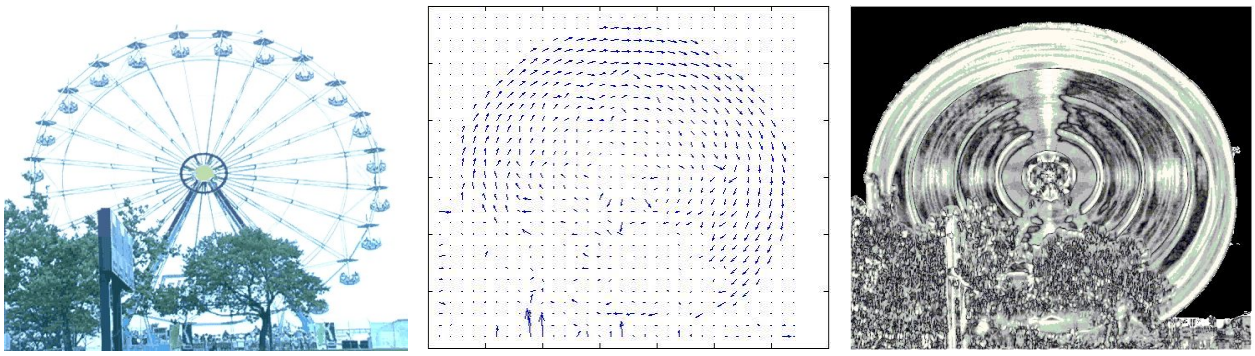
with a weighting factor $\beta_i = \alpha P(\mathcal{N}_i | \tilde{\nabla} I)$, which combines the probability that $\mathcal{N}_i$ is the correct model, with the factor α chosen as earlier according to the desired adaptivity. However, if the maximum likelihood model is $\mathcal{M}_{unk}$, there is a strong probability that the image motion does not come from any of the current models, and so we use this measurment to initialize a new tensor field, $\mathcal{N}_{M+1}(0, \Sigma_{M+1})$. The complete update of the adaptive mixture model requires that the weights of the components be adjusted. The weights w_i can be updated as $w_{i,t} = (1 - \beta_i) w_{i,t-1} + \beta_i$.

The constraint on the derivative measurements at each pixel represented by the structure tensor is independent of the measurements at other pixels, and the block-diagonal form of each of the components of the mixture model maintains this independence. The mixture model implies that all measurements at a given time in the image come from one of the components. Let $W_i(t)$ be the event "the motion in



(a) Video (b) Estimated Flow (c) Confidence

Figure 1: Flow estimated from 3-D structure tensor for 10 minutes of video of an intersection. Multiple motions at the middle of the intersection cause a circular pattern in the estimated flow field. The confidence decreases in the parts of the middle of the intersection where the angle between the two major motions is larger.



(a) Video (b) Estimated Flow (c) Confidence

Figure 2: **Flow estimated for a moving Ferris wheel. Pixels that have not experienced much motion (e.g., the lower left of this image) may have nearly rank-deficient Σ, resulting in falsely elevated S values.**

the world comes from model i at time t". Then for pixels $p, p', p \neq p'$, our covariance constraint can be rewritten as

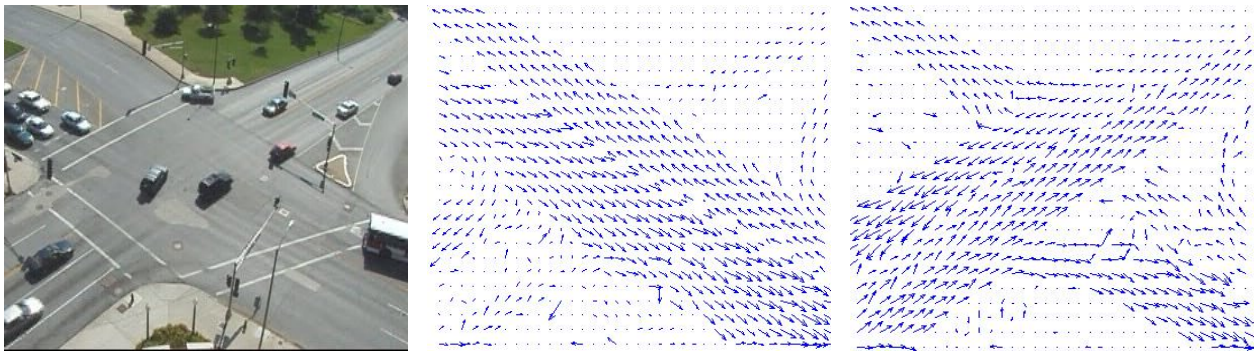
$$P_{p,p'}(\nabla I_p, \nabla I_{p'}|W_i(t)) = P_p(\nabla I_p|W_i(t))P_{p'}(\nabla I'_p|W_i(t)).$$

That is, measurements at different pixels are conditionally independent, given that motion in the world comes from model i. Using this choice of a global model to express all of our knowledge about inter-pixel dependencies allows the model to be maintained efficiently. One final note, because the motion fields are generated by discrete objects, in no frame is the entire component motion field visible, even if single frame optic flow measurements were reliable, it would not be possible to generate these components (with a standard EM type approach).

3.3 Obtaining a Meaningful Clustering

Figure 3 shows the mixture model estimated by the simple adaptive mixtures aglorithm described in the previous section. The algorithm clearly discovers the two major modes of this scene. However, finer features such as cars turning left are lost in the clustering process. The main difficulty in producing a clean segmentation is that while flow fields are defined over the entire scene, at any given frame there is unlikely to be motion everywhere. This leads to difficulties in bootstrapping and initializing new models.

We address this problem by considering sequences of coherent motion within the video, and assigning these sequences to a single model. Suppose measurements $A = \{\tilde{\nabla} I_{t-L}, \tilde{\nabla} I_{t-L+1}, \ldots, \tilde{\nabla} I_{t-1}\}$ have already been judged to come from a single motion. We can determine whether the next measurement, $\tilde{\nabla} I_t$, comes from the same discrete mode aggregating the measurements A into a single Gaussian $\mathcal{N}_{new}(0, \Sigma_{new})$. Then, $\tilde{\nabla} I_t$ is judged to belong to the same discrete motion if $P(\tilde{\nabla} I_t|\mathcal{N}_{new}) > P(\tilde{\nabla} I_t|\mathcal{M}_{unk})$. If $\tilde{\nabla} I_t$ is judged to come from $\mathcal{N}_{new}$, we use it to update $\mathcal{N}_{new}$. Otherwise, we initialize a new Gaussian $\mathcal{N}'_{new} = \mathcal{N}(0, \tilde{\nabla} I \tilde{\nabla} I^T)$ and assign A to one of the mixture components, $\mathcal{N}_1, \ldots, \mathcal{N}_M$.



(a) Intersection (b) $\mathcal{N}_1$ (c) $\mathcal{N}_2$

Figure 3: The two highest-weighted mixture components found by the algorithm of section 3.2. The two major motions in the center of the intersection are clearly differentiated.

Treating frames as independent, the negative log-likelihood $-\log P(A|\mathcal{N}_i)$ is just $\sum_{i=t-L}^{t-1} -\log P(\tilde{\nabla} I_i|\mathcal{N}_i)$. The posteriors $P(\mathcal{N}_i|A)$ can then be calculated as in the previous section. All of A can be assigned to the mixture component that maximizes the posterior or used to initialize a new mixture component, if $\mathcal{M}_{unk}$ is the maximum aposteriori mixture component. Let $\mathcal{N}_j(0, \tilde{Sigma}_j)$ be the best mixture component and α the weight given to each new frame, as in Section 2.1. We can update $\mathcal{N}_j$ wholesale as $\tilde{\Sigma}'_j = \gamma\tilde{\Sigma}_j + (1-\gamma)\tilde{\Sigma}_{new}$. Setting $\gamma = (1-\alpha)^L$ preserves the exponential decay process by producing the same weightings as if each frame was added sequentially. Since $\tilde{\Sigma}_j$ can be updated directly from covariance $\tilde{\Sigma}_{new}$, it is not necessary to keep every $\tilde{\nabla} I_i \in A$ in memory.

4 Results

Because the algorithm only considers pixels with nonzero I_t, the computational cost is somewhat data-dependent. However, on a 2.3 GHz Pentium 4 PC, our C++ implementation achieves real-time performance (20 fps at 360x240 resolution) on every dataset we have applied it to. When run on video of a water scene with motion at almost every pixel, it maintained up to 6 tensor fields, a rough lower bound on the real-time capability of the algorithm. The "Intersection" dataset is more typical of actual surveillance applications and contains sparser motion. On this dataset, our system can maintain up to 10 tensor fields in real time, far greater than the actual number of modes of motion in the scene. The only specific optimization applied in generating these results was to amortize the cost of matrix inversion by only recalculating Σ^{-1} every 10 frames.

The main free parameter in our algorithm is the prior model for foreground motion, or the choice of threshold if the prior model is a uniform distribution. In scenes with sparse motion, the segmentation of gross features such as the two principal directions of motion in the intersection is relatively insensitive to the specifics of the foreground prior. However, finer features such as motions that occur in only a small part of the image or motions that occur infrequently require some hand tuning of parameters. The result in Figure 3 was generated using a foreground prior realized as a threshold mahalanobis distance of 17 per moving pixel. The result in Figure 4 was generated by the block-clustering approach of Section 3.3, using an Gaussian prior, $\mathcal{N}(0, 80I)$, at each pixel.

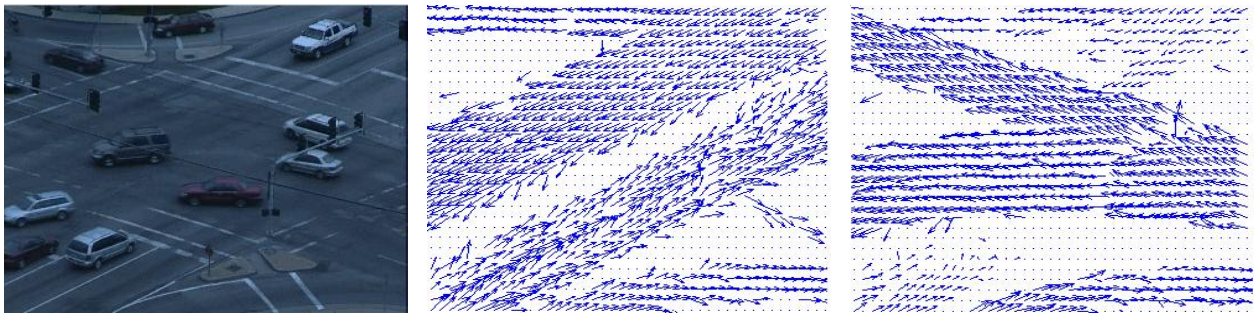
In practice, the image derivatives are calculated using Sobel filters applied to Gaussian (spatially) blurred images, and differences between image to calculate I_t. This causes the magnitude of noise in I_t to be far greater than that in I_x and I_y. The derivation of the third eigenvector as the maximum-likelihood optic flow assumes white noise in I_x, I_y, I_t. In practice, we have found that computing optic flow from the matrix equation:

$$\begin{pmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = - \begin{pmatrix} \sum I_x I_t \\ \sum I_y I_t \end{pmatrix}$$

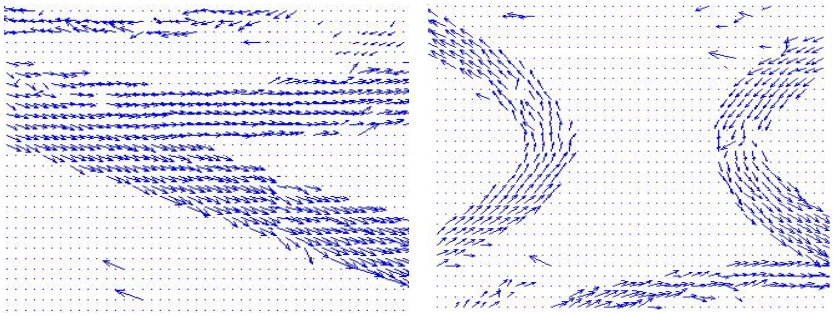
provides a more robust estimate than the third-eigenvector solution. All plots in this paper have been generated in this manner.

5 Conclusion

We have demonstrated the use of the structure tensor field as a model of background motion. We have shown how a set of tensor fields can be viewed as a single Gaussian mixture model, leading to compact and computationally efficient representation of inter-pixel dependencies. The algorithm runs in real time and can adapt to both slow drift and abrupt changes in motion patterns. Because it considers covariances between filter responses, rather than intensity it-



(a) Intersection (b) $\mathcal{N}_1$ (c) $\mathcal{N}_2$



(d) $\mathcal{N}_3$ (e) $\mathcal{N}_4$

Figure 4: Four mixture components estimated using the block clustering approach of section 3.3.

self, the algorithm is somewhat robust to lighting changes in the scene. The factoring of image motion into component fields is important for potential applications in anomaly detection and tracking.

References

[1] Bigun, J. and Granlund, G. H., and Wiklund, J. "Multidimensional orientation estimation with applications to texture analysis and optical flow.", *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 13 (8): 775-790, 1991.

[2] Martin Grger, Wolfgang Sepp, Tobias Ortmaier and Gerd Hirzinger. "Reconstruction of Image Structure in Presence of Specular Reflections", In *Proceedings of the 23rd DAGM Symposium on Pattern Recognition (DAGM 2001)*, pp 53-60, 2001.

[3] Haritaoglu, I. and Harwood, D. and Davis, L.S., "W4: A Real Time System for Detecting and Tracking People", *CVPR98*, pp. 962-962, 1998.

[4] Horn, B.K.P. and Schunck, B.G., "Determining Optical Flow", *MIT AI Memo*, 1980.

[5] Mester, R., "A new view at differential and tensor-based motion estimation schemes", *Pattern Recognition 2003, Proceedings of the 25th annual conference of the Deutsche Arbeitsgemeinschaft fr Mustererkennung (DAGM)*, Magdeburg, Germany, 10-12. September 2003.

[6] Mittal, A. and Paragios, N. "Motion-Based background subtraction using adaptive kernel density estimation", *CVPR 2004*, 2004.

[7] Monnet, A., Mittal, A., Paragios N. and Ramesh, V. "Background modeling and subtraction of dynamic scenes", *Proc., IEEE Int'l. Conf. on Computer Vision (ICCV '03)*, pp. 1305-1312. 2003.

[8] Pless, R., and Jurgens, D., "Road extraction from motion cues in aerial video", *Proceedings of the ACM-GIS Workshop*, 2004.

[9] Pless, R., Larson, J., Siebers, S., and Westover, B., "Evaluation of Local Models of Dynamic Backgrounds", *CVPR 2003*, 2003.

[10] Stauffer, C. and Grimson, W.E.L. "Adaptive Background Mixture Models for Real-time Tracking", *CVPR99*, Vol. II, pp. 246-252. 1999

[11] Zhong, J. and Sclaroff, S. "Segmenting foreground objects from a dynamic textured background via a robust Kalman filter", *ICCV03*, pp. 44-50, 2003.