of the interaction terms are smaller than the diagonal elements and are ignored. The partial derivative within a single equation however do not have the same noise distribution due to the elongated Gaussian filters used (.i.e. $< I_x^2 > \neq < I_y^2 >$). The partial derivative along the orientation of the filter has a higher noise response than the one perpendicular to the orientation. Since the oriented filters come in rotated pairs, a simple sum and difference of the two resulting equations produces two equations with partials of equivalent noise variance yet remain independent. The total least squares solution (7) is simply

$$\vec{\mathbf{v}} = \left(\begin{bmatrix} \sum I_{xi}^2 \phi_i & \sum I_{xi} I_{yi} \phi_i \\ \sum I_{xi} I_{yi} \phi_i & \sum I_{yi}^2 \phi_i \end{bmatrix} - \sigma_3^2 \mathbf{I} \right)^{-1} \begin{pmatrix} -\sum I_{xi} I_{ti} \phi_i \\ -\sum I_{yi} I_{ti} \phi_i \end{pmatrix} \tag{14}$$

The summations are over the filtered responses within that scale group. The weights ϕ_i are the expected variances of each measurement. This formula is very similar to the one used by many authors [SAH91, CV90, etc.] who attempt to overcome the underconstrained system (2) by making an implicit smoothness assumption. These approaches sum over the responses from surrounding pixels, thus assuming the velocity to be the same within a local region.

Notice that we need only invert a 2x2 matrix whose entries are simply weighted sums of filter outputs. Thus only simple operations are required and can be performed in parallel. The filter responses can be accumulated as they are produced and need not be stored in memory. Since the measurement matrix has rank 3, simple explicit formulas exist for the three singular values, σ_i. They come from solving a cubic equation whose coefficients are combinations of the summations which appear in (14). Since singular values are always real and non-negative, a simpler form of the general solution of the cubic equation can be used. Testing the condition number of the matrix is then just the ratio of the two largest singular values. If the condition number is above 100 the estimate is discarded. This rarely occurred in our simulations. Each scale group now evaluates its velocity estimate. The magnitude of this velocity estimate is compared with the maximum magnitude allowed for this scale group and rejected if larger. The ratio of the smallest singular values is compared with the consistency threshold.

The scale estimates that are not discarded are combined in a final, multi-scale total least squares framework. The weighting terms for this combination is modified by scaling the weight from the first matrix by the inverse of the relative error term. Thus the diagonal matrix elements are replaced by:

$$\phi_i = \begin{cases} \frac{\phi_i}{\sigma_3/\sigma_2 + \epsilon} & \sigma_3/\sigma_2 < C_t \text{ and } |\vec{\mathbf{v}}| < \mathbf{v}_{max} \\ 0 & \text{else} \end{cases}$$

where ϵ is a small number to prevent division by zero. This weighting gives more credit to scale groups with estimates which best match the constraint equations.

The final estimate computed by combining scales is rejected if the ratio of singular values indicates the equations are deemed inconsistent according to the consistency threshold. One place where this can happen is if different spatial scales are seeing different motions due to some kind of motion boundary. Our future work will look at ways of identifying and resolving these situations. For now, they are simply labeled as places without estimates (holes).

5 Experimental Results

For the experiments where the true optical flow is known, we use the angular error measure used by Barron et al. [BFBB92] to evaluate the results. They measure the error between the true velocity $v = (v_1, v_2)$ and the estimate $\hat{v} = (\hat{v}_1, \hat{v}_2)$ as the angle between the unit vectors in 3 space, $\vec{\mathbf{v}} = (|v|^2 + 1)^{-1/2}(v_1, v_2, 1)$.

$$\psi_\epsilon = \arccos(\vec{\mathbf{v}} \cdot \hat{\mathbf{v}}) \tag{15}$$

This is calculated at every pixel value where an estimate is formulated. Also reported are the percentage of pixels without estimates (holes).

5.1 Synthetic Data

Systematic Error. A random dot pattern with variance 1000 units is translated diagonally. Eack pixel was uncorrelated, thus all frequencies were present. Figure 5 shows the angular error and percentage of holes (no estimate)

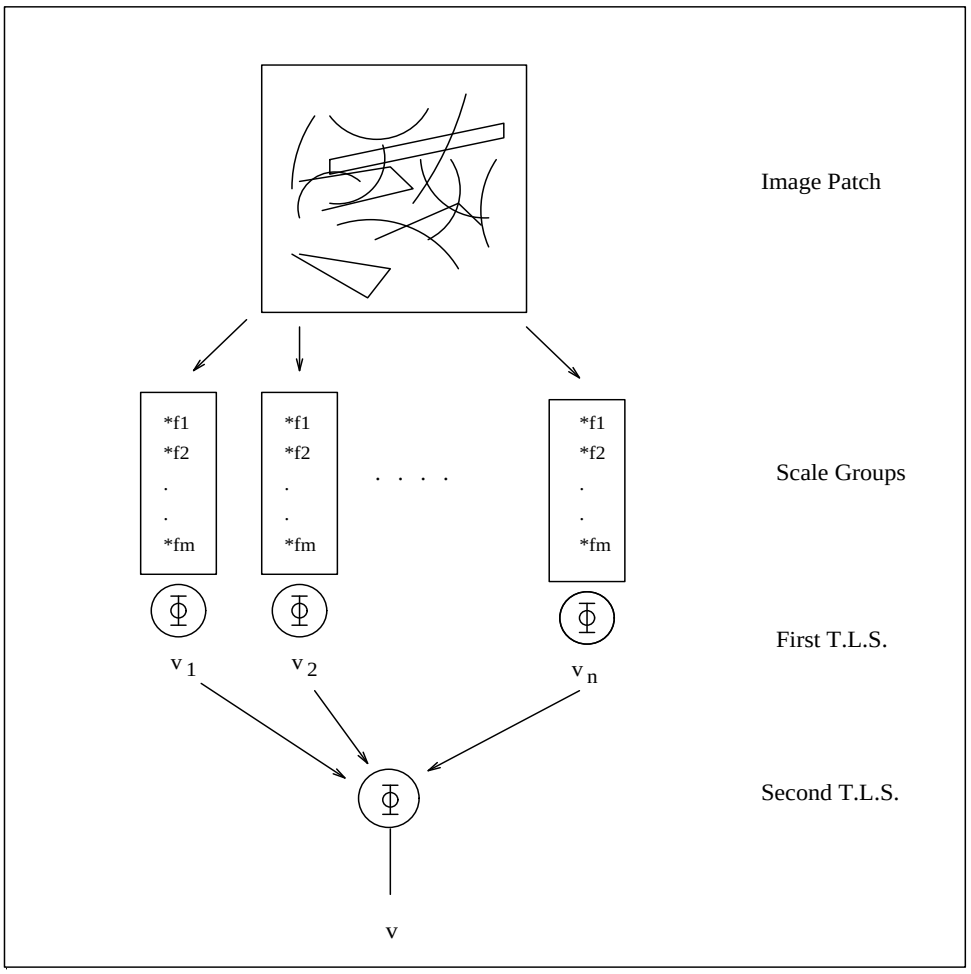

Figure 3: Multi-scale gradient technique for motion. A patch of the image is convolved with groups of linear spatio-temporal filters. Each group is tuned to the same spatial scale. Each group makes its own estimate for the velocity using total least squares. The estimates are combined in a second re-weighted total least-squares formulation. The magnitude of the velocity estimate a group may present is limited by the systematic error.

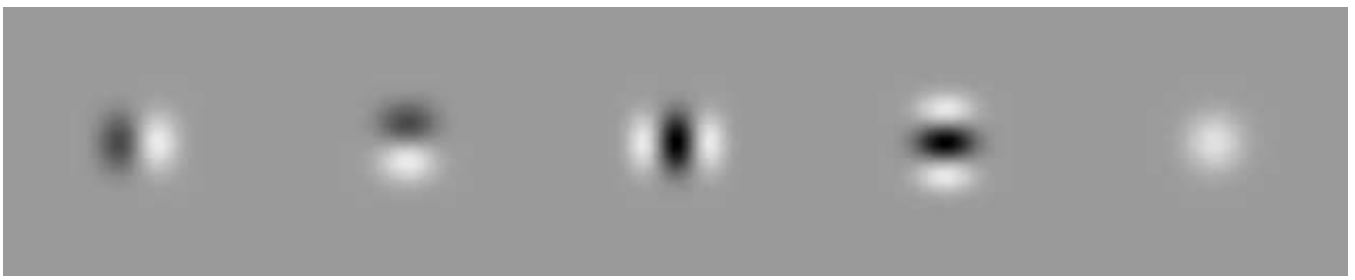

Figure 4: Spatial impulse response of the filter set used for a single spatial scale. The filters are either the Laplacian of a Gaussian or products of Gaussians and a Gaussian derivative. These filters all have a zero DC response. The same set of filters is used in some approaches to early vision.

equation (13). The estimate is also rejected if the ratio of the two smallest singular values of the measurement matrix is above the consistency threshold, C_t.

The remaining valid estimates are combined again via a total least squares formulation. The weights in this step have been divided by the consistency estimate of that scale's equations. This was the ratio of the two smallest singular values of the measurement matrix.

This two step method contains two elements which make it robust. First, the ratio threshold prevents scale groups with poor estimates from participating in the second stage. Secondly, those scales which do participate are weighted by their individual residuals. In many iterative robust techniques, the process of finding an estimate, weighing by updated covariances and repeating is common [Hub81].

By weighting the second stage by the singular values ratio, we insure that the scale which most accurately estimates the motion has the strongest influence in the multi-scale fusion. This is in contrast to coarse-to-fine methods which assume larger scales have a correct but coarse estimate.

A consistency ratio for the combined scales is also available. This serves as a useful indicator of accurate optical flow. If this ratio is larger than the threshold C_t that pixel location is marked as not having an estimate. We feel that this may indicate the presence of motion boundaries and our future work will investigate this. Figure 3 outlines the method.

4 Implementation

The input to the model is simply two response frames separated in time. These two inputs are created from a sequence of images via convolution with separable Gaussian functions. First, each frame of the sequence is convolved with spatial functions. The functions used were the standard normalized Gaussian function, edge and bar filters which consists of a derivative of a Gaussian along one spatial dimension and a standard Gaussian along the orthogonal spatial dimension, and the symmetric Laplacian of a Gaussian. These filters have been used to model receptive fields of neurons in the visual cortex [You85]. It has been argued that they form a good basis for early vision tasks such as edge detection, stereo and texture discrimination [Koe88, JM92b, MP90]. The size of each Gaussian function was set so that each filter shared a common peak frequency response. Thus they shared a common scale, σ_0. Figure 4 is the spatial responses for the scale group of $\sigma_0 = 16$ pixels for the first 2 Gaussian derivatives and the Laplacian of a Gaussian. This set was the one most often used in the experiments. Notice that each filter has a zero DC response and thus is not influenced by global lighting changes.

The sizes of the scale groups followed the progression $\sigma_0 = 1, \sigma, \sigma^2, \ldots, \sigma^{n-1}$ where σ was 1.8 for the experiments. This is a natural scale space representation as used in pyramid implementations. A pyramid scheme can be used to decrease the computational load for the many convolutions required without a significant loss in performance (see Section 5.3).

Next, the filtered responses are convolved in the time dimension with a causal standard Gaussian. This is non-zero only for past time frames. The standard deviation of this Gaussian was set to 3 video frames in order to emulate human response curves which show temporal recruitment up to about 100 milliseconds. Thus only the past 10 frames contribute significantly to any filter response. These numbers could change depending on the frame rate or known motions.

The partial derivatives of the responses are computed through a forward finite difference cube. This is simply the average of 4 adjacent pixel forward finite differences. Since the stimulus was already convolved with Gaussian functions, we felt a more sophisticated scheme for obtaining first partials was not needed. The forward differences actually provide an estimate of the partial derivatives on a lattice which is offset one half pixel in each spatial dimension and one half of a frame in the temporal dimension.

If a given scale group contains adequate texture such that the condition number of the measurement matrix was finite, a velocity estimate for that scale group is computed. It is known that it is important to 'whiten' the measurement matrix such that each element is identically distributed and independent [VV91]. We assume that the output of each filter is independent. For the zero-DC filters used, the oriented filters within a scale group actually are orthogonal. The Laplacian is not a linear combination of the elongated second order derivatives, but it is not orthogonal. When the different scale groups are combined, the filters are not orthogonal, but the magnitudes

find that the estimated velocity, $\hat{v}$, as a function of the true velocity, v, is

$$\hat{v} = \sin(\omega v)/\sin(\omega) \tag{10}$$

The difference between v and $\hat{v}$ comes from the linear approximation of finite differences. The relative error for this component is

$$|\frac{\hat{v}-v}{v}| = 1 - \frac{\sin(\omega v)}{v\sin(\omega)} \tag{11}$$

Writing this as a function of β at the maximum velocity,

$$|\frac{\hat{v}-v}{v}| = 1 - \frac{\omega\sin(2\pi\beta)}{2\pi\beta\sin(\omega)} \tag{12}$$

We find for a range of reasonable frequencies that a value of β about $1/2\pi$ results in a fractional error of less than 10%. Thus we set $\beta = 1/2\pi$. The resulting limit on the displacements allowed is $|v| < 1/\omega$.

Unfortunately we do not know the stimulus of the intensity function. However in our implementation, the intensity function is convolved with a series of Gaussian and Gaussian derivative functions. These are either low or band-pass filters. Thus we know the expected frequency response of the stimulus. If we use a low pass filter with a cutoff frequency ω_c then the maximum velocity estimate which can be considered valid is $|\vec{\mathbf{v}}| < 1/\omega_c$ from the above analysis. The filters we use consist of Gaussian functions and their derivatives. The n'th derivative of a Gaussian of standard deviation σ has its maximum frequency response at $\omega = \sqrt{n}/\sigma$ [You85]. If we use this frequency in limiting the maximum displacement we find that for the response formed by filtering with the n'th derivative of a Gaussian, the maximum displacement we can accept without knowing the stimulus spectrum is

$$|\vec{\mathbf{v}}| < \sigma/\sqrt{n} \tag{13}$$

2.3 Systematic errors due to model failure

There are situations where the underlying assumptions of the model are violated. Constancy of image brightness (1) is not strictly true whenever there is a significant specular component or when occlusion occurs. The model also assumes that the optical flow field is locally constant over the support of the filters, which is not true if the filter support straddles a depth discontinuity or when there is a significant rotational or dilational component in the flow field. In these situations, regression is not valid and these measurements should be labeled as outliers.

When calculating the total least squares solution, the singular values of the measurement matrix are available. The smallest singular value, σ_3, is equivalent to the Frobenius norm of the perturbation needed to make the equations consistent. We define a consistency ratio $\frac{\sigma_3}{\sigma_2}$. This is the ratio of the norm of the perturbation to the smallest eigenvector of the resulting measurement matrix. When the assumptions are violated, a large relative perturbation will be needed to make the equations consistent. We discard scale groups which require a perturbation greater than a given threshold, C_t to become consistent, assuming they are due to model failure. This discards the outliers before combining scales in a second total least squares. The ratio is scale independent and therefore can be used to compare estimates between scales.

3 Model

Based on the above analysis, our model for multi-scale motion analysis consists of two total least squares steps. The image is first convolved with a collection of filters. These filters are separated according to scale. Thus we may have m different filters each of the same scale but differing in orientation and frequency response, and n such groups of these filters. The common scales of these groups form a geometric sequence. If the smallest scale is of size σ_0, then the i'th scale is of size $\sigma_o^{(i-1)}$.

Partial derivatives are computed via finite differences and weighted according to known prior noise variances.

The n scale groups each individually form an estimate for the velocity via the total least squares formula in (7). This velocity estimate is deemed valid if its magnitude is less than the maximum allowed for that scale via