Image enhancement and moving target detection in IR image sequences

Wouter Beck

TNO Physics and Electronics Laboratory
P.O. Box 96864, 2509 JG The Hague
The Netherlands
e-mail: wbeck@fel.tno.nl

ABSTRACT

Results are presented of noise reduction by motion compensated temporal filtering in a noisy IR image sequence and of moving target detection in an air-to-ground IR image sequence. In the case of motion compensated temporal filtering our approach consists of estimating the optical flow between successive frames and subsequently averaging a small number of images. Moving targets are detected by first estimating the optical flow between successive frames. Target detection amounts to comparing a predicted frame, based on the estimated optical flow, to the actual frame. Thus, it is possible to detect targets without making assumptions on their appearance. The particular motion estimator used was found to be especially useful in the case of IR imagery, because the estimator is relatively insensitive to noise and global brightness variations.

1 INTRODUCTION

Motion perception is an important source of information for the human visual system. The determination of our motion relative to the environment as well as the determination of the three dimensional structure of the environment largely depend on the interpretation of visual motion. The human visual system is capable of extracting information from a sequence of images that is hard to extract from the individual images. An example is the interpretation of a very noisy image sequence. By using spatial and temporal correlation we are able to "see through the noise." Sometimes, visual detection of an object fully depends on the perception of motion. This is illustrated by the ease with which we see an otherwise successfully camouflaged object as soon as it moves.

2 IMAGE MOTION ESTIMATION

The apparent motion of brightness patterns observed when a camera is moving relative to the objects being imaged is called optical flow. Optical flow can be represented by a two dimensional vector field.

0-8194-1269-4/93/$6.00

Loosely speaking, the optical flow field links a pixel at the position (x, y) to the corresponding pixel at position $(x + u(x, y), y + v(x, y))$ in the next image. Ideally, both pixels correspond to the same physical object point in the scene. In practice, when estimating optical flow, this is hard to achieve because there is an infinite number of vector fields that is consistent with the data. There is a large body of literature devoted to the various approaches of estimating optical flow from a sequence of digitized images.

The particular approach to the motion estimation problem we have taken is described in more detail by Beck[1] and is based on concepts introduced by Jepson and Fleet.[2] One of the central ideas of our motion estimator is the separation of the analysis according to scale and orientation. A *multiresolution representation* provides a simple hierarchical framework for analyzing the image information. The primary reason for using a multi resolution representation is not computational efficiency. The concept of multi resolution is deemed to be essential for the task at hand, because the information in images resides at different resolutions and needs to be analyzed at the appropriate resolution. Large scale image structure provides the context in which smaller scale structure fits. In motion estimation, large scale image structure is used to measure large displacements with limited accuracy. These coarse measurements are subsequently improved by using information provided by the finer scale image information. This avoids aliasing problems and at the same time avoids a temporal sampling rate that would be prohibitively high. Prominent structural image information can usually be characterized by a specific orientation. Performing measurements on image structure with primitives that do not have approximately the same orientation is bound to produce ill-conditioned results. Therefore, we use a multi resolution representation with basis functions that are all rotations of one unique function.

The most commonly chosen image attribute for motion estimation is brightness. Usually, the main assumption is that a pattern of image brightness moves across the image plane without distortion. In practice, this assumption is often violated. Therefore it would be preferable to define an image attribute that more inherently describes image structure. For this purpose, Jepson and Fleet[2] proposed the use of local phase information from a pair of quadrature bandpass filters. It turns out that phase has several desirable properties. Phase is amplitude invariant, and hence insensitive to global variations of image intensity. Because phase is computed from information in a (small) neighborhood, noise sensitivity is reduced by the implied averaging. Another way to view this reduced noise sensitivity is the fact that generally the noise spectrum extends over the entire frequency plane. Since phase is computed from the output of a bandpass filter, it is only sensitive to that part of the noise that falls within the passband of the filter.

3 NOISE REDUCTION IN IMAGE SEQUENCES

By $I(\mathbf{x}, t)$ we denote the image brightness function, where vector $\mathbf{x}$ denotes the spatial coordinates and t denotes time. Let $v(\mathbf{x}, t)$ be the displacement of the image point at $(\mathbf{x}, t)$ between time $t - \Delta t$ and t, where Δt denotes the temporal sampling interval. Assuming that image brightness for an object point is conserved over time, we can write

$$I(\mathbf{x}, t) = I(\mathbf{x} - v(\mathbf{x}, t), t - \Delta t) \tag{1}$$

Obviously, v is undefined when an object is occluded or when it is newly exposed. In general v will be a slowly varying function of the spatial coordinates with discontinuities at the edges of moving objects. A spatiotemporal volume can be formed by stacking the consecutive frames of the sequence. A physical

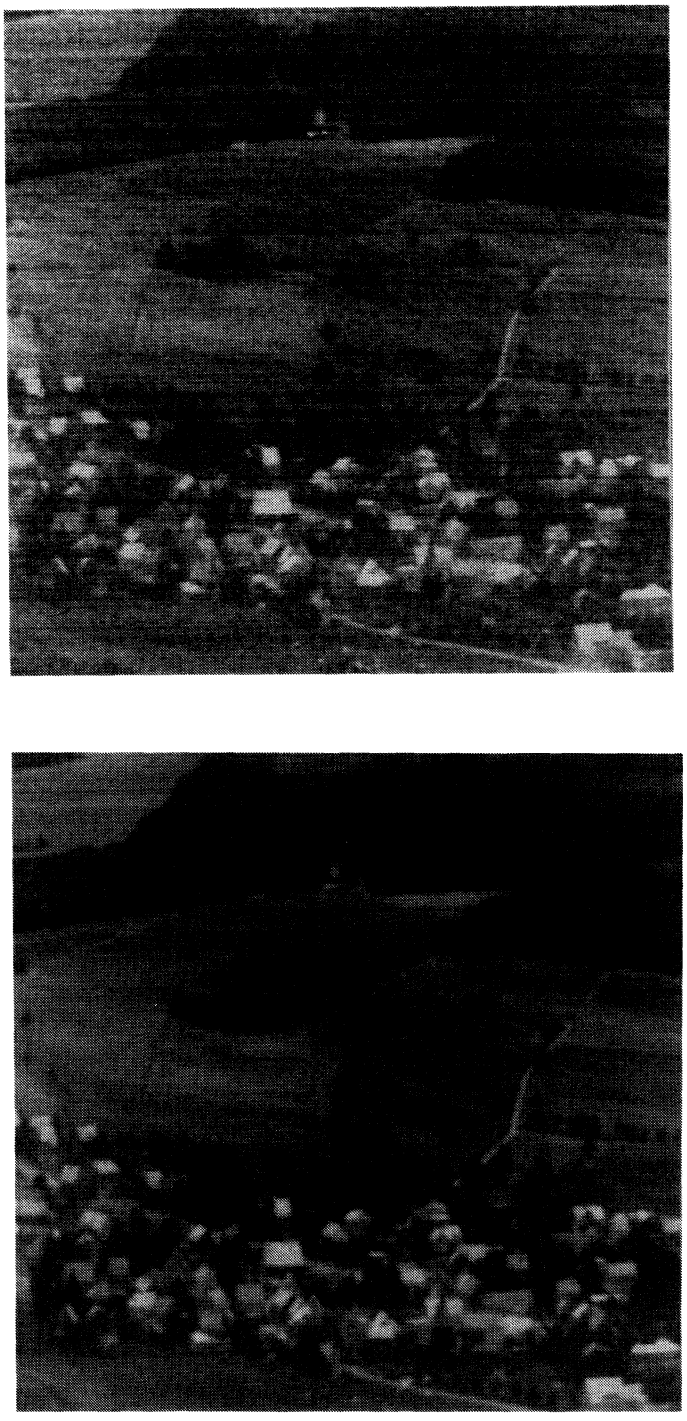

Figure 1: (Top) Part of the reference image from an IR image sequence, taken from an aircraft flying at approximately 85 m/s at a height of about 1000 ft. Frame rate 8.3 Hz. All computations are performed relative to this image.
(Bottom) Mean along motion trajectories computed from six motion compensated images and the reference image. Median filtering along motion trajectories produces similar, though generally slightly noisier, results.

point in the scene traces out a trajectory in this spatiotemporal volume during the time it is visible in the sequence. The brightness value along this trajectory forms a one dimensional signal. This signal is assumed to consist of a deterministic image component and an additive noise component. Variation of the image component is due to change in the luminance of the object. This variation is assumed to be relatively slow, so that the image component is a low bandwidth signal. The additive noise is assumed to be uncorrelated with the image signal. Low-pass filtering along the motion trajectory can significantly reduce the noise component. The filter operation along the motion trajectory can be either linear or non-linear. When the image noise is additive Gaussian noise, independent in each pixel and of fixed variance along a motion trajectory, then it can be shown that the sample mean along a motion trajectory is the maximum likelihood estimator for the grey value of the pixel. In this case, the linear estimator will yield the best signal to noise ratio in the result. On the other hand, a non-linear filter may be more robust to errors in the displacement estimate and the non-validity of the noise model e.g. in the case of dead pixels in the images. In addition, a non-linear filter might be able to deal with occlusion and exposure effects more adequately. The choice of filter will generally depend on the ease of implementation and the particular distortions in the image sequence. The number of frame stores can be reduced if the used filter is recursive.[4]

With regard to exposure and occlusion effects, it would be of interest to know exactly the lifetime of a motion trajectory. Unfortunately this is a hard problem. It requires the identification of image areas that are newly exposed and image areas that are just occluded in each frame of the image sequence. Most current motion estimators are not able to solve this problem reliably.

4 DETECTION OF MOVING TARGETS

Algorithms for the detection of dim, low contrast targets usually consist of two stages. First, the algorithm selects a number of potential targets, for example bright spots. Due to clutter this usually results in a large number of false alarms. (Here, clutter is loosely defined as the amount of target-like objects in a scene.) The second stage therefore has to reject the falsely selected objects. This can be done by combining information over frames, or by use of contextual information. Here, we choose to detect targets on basis of their motion. Our approach consists of essentially two stages:

1. motion estimation,
2. target detection in the motion compensated image sequence.

We assume we have to deal with an essentially stationary scene that is being imaged from a moving platform (e.g. helicopter). If we are able to estimate a sufficiently accurate 2-D vector field that maps one frame in the sequence to the next, we can in principle predict one frame from the previous one. The principle to detect moving targets is particularly simple and amounts to analyzing the image sequence on the occurrence of unexpected events. In this context, unexpected events are temporal variations of the image brightness function that are impossible to predict and that can not be accounted for by noise. Thus, targets are detected by analyzing the difference between the predicted and the actual image. In principle, it is possible to detect camouflaged targets moving relative to a textured background.

In evaluating the difference images thus obtained, we have to distinguish several possibilities.

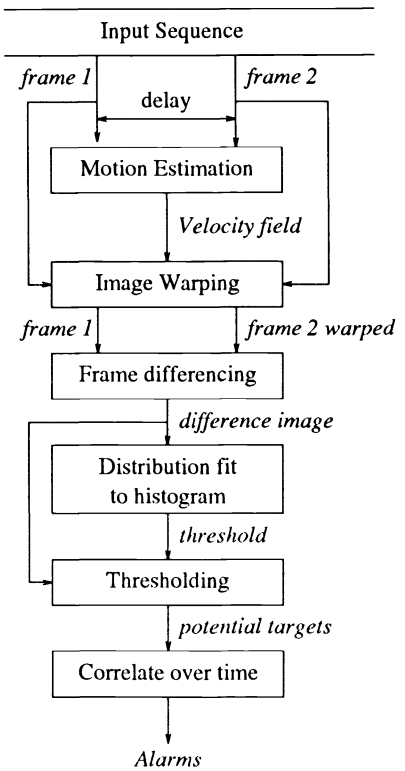

Figure 2: Block diagram for the detection of moving targets in an image sequence obtained by a moving camera.

1. When the image motion estimate is perfect, the scene has no appreciatable depth discontinuities, and there are no moving objects in the scene, we expect the difference image to be a sample from a 2-D random noise process. The noise is a mixture of image noise and noise due to the interpolation process.

2. When the motion estimate is accurate, there are no moving objects, but there is considerable depth variation in the scene, we expect uncovered background adjacent to physical edges of foreground objects. This is the parallax effect. Generally, this will result in large amplitudes in the difference image at locations corresponding to covered and uncovered background, while the rest of the difference image is characterized by random noise. In air-to-ground imagery the large response areas will usually be chain-like, for example the outline of a hill. Although the large response areas will generally not correspond to moving targets, they are nonetheless of interest because they often correspond to previously unexposed parts of the scene. In some applications it may be of interest to perform extra processing on parts of the scene that are newly exposed.

3. When there are moving objects in the scene and the motion estimate is such that this object motion is correctly captured, the difference image will show large amplitudes at locations of covered and un-covered background if the background is sufficiently textured. This enables us to detect camouflaged objects.

4. When there are small, moving objects we may be unable to capture object motion correctly. This behavior may be forced by only using large scale image structure in the motion estimate. In

this case the difference image will generally display a small area with a large positive response adjacent to a small area with a large negative response. This case is of practical interest in target acquisition applications at long stand off ranges.

To automate the detection process, we have to make a number of assumptions about the image noise statistics. In the examples shown here, the noise was assumed to be additive zero-mean Gaussian noise, independent in each pixel. The noise statistics are obtained from a fit of a Gaussian to the sample histogram of the difference image. This is more robust than calculating the usual sample statistics because the influence of the outliers (targets !) is reduced. From the standard deviation thus obtained, a statistically meaningful threshold may be obtained. The confidence in the presence of potential targets may be increased by correlating the detection results over time. This may involve more or less sophisticated techniques such as described by Blostein and Huang.[3] The overall procedure for detection of moving targets can be summarized by the scheme of figure 2.

4.1 Target detection using target motion

The upper photograph of figure 4 shows a frame from an air-to-ground IR image sequence. In this sequence there are several moving targets, cars on the roads. Notice that in this sequence the contrast is inversed, i.e. hot areas appear dark. For the present algorithm, this makes no difference. For the target detection we used three frames $f(t)$ at times t_{-1}, t_0 and t_1. The image motion between $f(t_0)$ and $f(t_{-1})$ and between $f(t_0)$ and $f(t_1)$ was estimated using a phase based motion estimator.[1] Because this image sequence is contaminated by a fair amount of noise and sensor artifacts, and because this image sequence lacks small scale image structure in certain parts of the scene, we used a planar patch model to locally improve the estimated image motion. It can be shown[5] that the planar patch model is described by the mapping:

$$x' = \frac{A_{11}x + A_{12}y + A_{13}}{A_{31}x + A_{32}y + 1} \tag{2}$$

$$y' = \frac{A_{21}x + A_{22}y + A_{23}}{A_{31}x + A_{32}y + 1} \tag{3}$$

Equations (2) and (3) define a mapping from the two-dimensional image-space (x, y) at time $t = t_1$ onto the image-space (x', y') at time $t = t_2$. The eight non-trivial parameters A_{ij} are the so called *pure parameters*. They are uniquely determined for a given motion and planar patch. The pure parameters are estimated from the image motion vectors produced by the phase based motion estimator. Both $f(t_{-1})$ and $f(t_1)$ are warped according to the estimated model (2) and (3) to obtain image estimates valid at $t = t_0$. These warped images are denoted by $\hat{f}_-$ and $\hat{f}_+$. First, we form the difference images $d_- = f_0 - \hat{f}_-$ and $d_+ = f_0 - \hat{f}_+$. Figure 3 shows a histogram of d_+. From figure 3 it is clear that this distribution is very well approximated by a Gaussian distribution, as shown by the dashed line. The parameters of this Gaussian were determined using a non-linear least squares fit to the histogram. Next, we apply a thresholding procedure to the difference images d_- and d_+. A threshold factor θ is selected. Let $d(x, y)$ be the pixel value at location (x, y) in either d_- or d_+, and let μ and σ be the corresponding mean and standard deviation, respectively, as determined by the histogram fit. We define a normalized d_n by

$$d_n(x, y) = \frac{d(x, y) - \mu}{\sigma}. \tag{4}$$

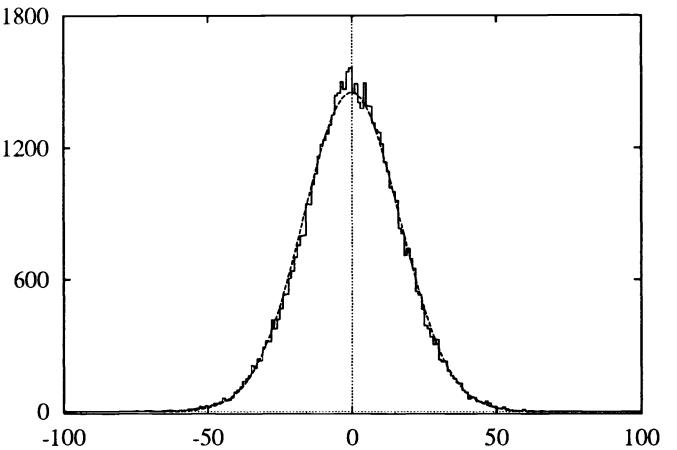

Figure 3: Histogram of the difference image d_+, obtained by subtracting the motion compensated image $\hat{f}_+$ from the reference image f_0. The dashed line represents the fitted Gaussian.

The result of thresholding procedure is determined by:

$$d_\theta(x,y) = \begin{cases} 0 & \text{if } |d_n(x,y)| < \theta/2 \\ \dfrac{2d_n(x,y)}{\theta} - \text{sign}(d_n(x,y)) & \text{if } \theta/2 < |d_n(x,y)| < \theta \\ \text{sign}(d_n(x,y)) & \text{if } |d_n(x,y)| > \theta \end{cases} \tag{5}$$

where $\text{sign}(\xi)$ is defined by

$$\text{sign}(\xi) = \begin{cases} -1 & \text{if } \xi < 0 \\ 0 & \text{if } \xi = 0 \\ 1 & \text{if } \xi > 0 \end{cases} \tag{6}$$

This procedure yields two frames of which almost all pixels are zero except for a number of positive and negative 'blobs ' with values between 0 and 1 and -1 and 0, respectively. This thresholding procedure has the advantage that it retains target responses that are not very strong. Of course, these 'weak' target responses have to be confirmed later on. Next, we discard all non-zero pixels in both frames that have opposite signs at corresponding positions. These 'cleaned' images are referred to as c_- and c_+. In the next step, we combine the images c_- and c_+ by pixel-wise multiplication. The positive blobs in the resulting image, referred to as T_0, correspond to potential targets.

5 REFERENCES

[1] W.J.C. Beck (1992), "Hierarchical computation of image velocity from local phase information," *Proc. 11th IAPR Int. Conf. on Pattern Recognition,* The Hague, pp. 526–529.

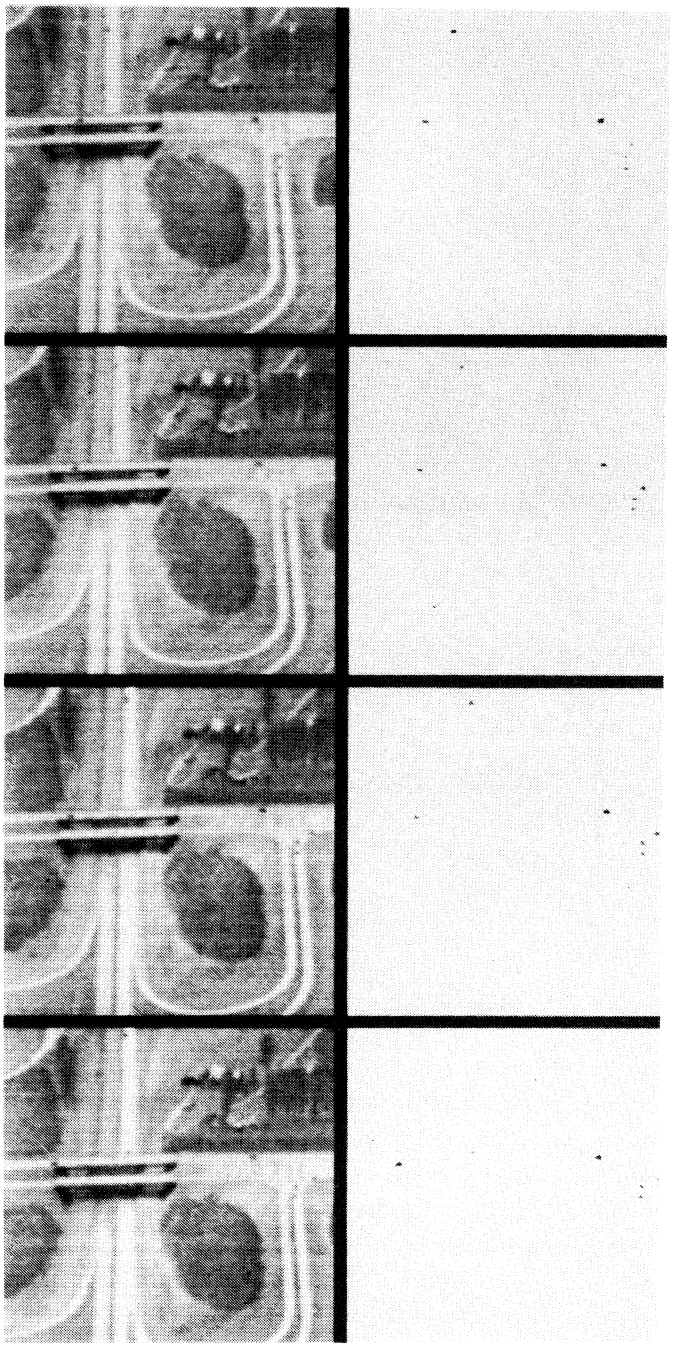

Figure 4: (Left) Four consecutive frames from an air-to-ground IR image sequence. (Right) Four consecutive frames with the detected moving targets. There are a few false alarms in individual frames. However, only the true targets are consistently detected. The false alarms could be eliminated by requiring consistency over time.

[2] David J. Fleet, Allan D. Jepson (1989), "Computation of normal velocity from local phase information," *Proc. Conference on Computer Vision and Pattern Recognition*, San Diego, USA, pp. 379-386.

[3] S.D. Blostein and T.S. Huang (1990), "Detecting small, moving objects in image sequences using sequential hypothesis testing, part i: Algorithms and Analysis," *IEEE Trans. on Acoustics Speech and Signal Processing,*

[4] E. Dubois and S. Sabri (1984), "Noise reduction in image sequences using motion compensated temporal filtering," *IEEE Trans. on Communications,* COM-32:826-831, 1984.

[5] R.Y. Tsai and T.S. Huang (1981), "Estimating three-dimensional motion parameters of a rigid planar patch." *IEEE Trans. on Acoustics Speech and Signal Processing,* ASSP-29:1147-1152, 1981.