Motion-Based Video Fusion Using Optical Flow Information

Jian Li, Stavri G Nikolov, Christopher P Benton and Nicholas E Scott-Samuel
University of Bristol, UK
{Jian.Li, Stavri.Nikolov, Chris.Benton, N.E.Scott-Samuel}@bristol.ac.uk

Abstract - Multi-sensor information fusion aims at extracting and combining useful information from different sensors. This paper addresses the problem of estimating and visualising motion information from a pair of visible and infrared cameras, using an optical flow technique. Videos from cameras sensitive to visible light are rich in texture and colour information such that a moving target can readily be positioned. On the other hand, videos from infrared cameras provide extra information which cannot be detected in the visible-light spectrum. In this paper we introduce a stochastic rule for combining optical flow computed from two (or more) sources. We also propose a novel motion-contingent selection method for the fusion of the co-registered visible and infrared video sources.

Keywords: optical flow, image fusion, motion information, motion fusion, least squares.

1 Introduction

In the task of recovering motion information from videos, gradient-based optical flow techniques have been widely used. These techniques are generally based on the Lucas-Kanade algorithm [1] and the Horn-Schunck algorithm [2]. The difference between these two methods lies in the different assumptions of the smoothness in the motion field. However, both techniques rely on the extraction of temporal derivatives of image brightness as the basis for their subsequent calculations.

Estimating motion information, such as velocity, relies on the successful detection of moving patterns from the video. Ideally, under the assumptions that there is no influence of noise and that the light source does not change, those positions corresponding to the non-zero temporal derivatives reveal the possible motion patterns. However, as the derivatives are solely computed from the illuminates, which are largely influenced by the imaging environment, the sensor property and the light conditions, derivative-based motion detection procedures can be compromised when using a standard visible camera. In Figure 1 (a), we show a single frame from a video captured from a visible (VIS) camera. We convert images into greyscale and filter them with the recursive temporal filter designed by Fleet and Langley [3]. The resultant image of the temporal derivatives in Figure 1 (b) (zero values are encoded by grey mean luminance) shows that only the motion of the walking person and the leaves can be detected by the visible camera.

![VIS video](a) VIS video ![VIS temporal derivative](b) VIS temporal derivative

![IR video](c) IR video ![IR temporal derivative](d) IR temporal derivative

Figure 1: Images from a multi-sensor system.

When we repeat the same procedure using a registered video captured from an infrared (IR) camera, the IR image shown in Figure 1 (c) illustrates that another man, who does not appear in the visible image, exists in the shadow. Correspondingly, in Figure 1 (d) the temporal derivatives in the body of the second man are non-zero. This information clearly illustrates the possibility that there is a second moving person in the same scene.

Compared to the visible image, the infrared image has the advantage of improved object detection in low-light conditions and provides illuminance invariance. The combination of infrared and visible imagery has been used in many applications in the past, such as facial recognition [4] and target tracking [5]. However, motion-based IR and VIS video fusion has rarely been investigated. As the above example shows, motion information derived from both IR and VIS sources is not necessarily the same, and both can be informative.
In this paper, we present a novel multi-source video fusion framework based on motion information. In this framework, videos are captured from a multisensor system, in which a VIS camera and an IR camera are placed side-by-side and fixed to the same tripod with a multicamera bracket. This setup allows the two cameras to record the patterns in the same scene so that image fusion methods can be readily applied. VIS and IR videos from such a system contain different information. On the one hand, the infrared video allows us to estimate the motion of hot moving objects in darkness or shadows. On the other hand, the visible video is rich in texture and colour information so that the motion information can be interpreted easily and be located spatially. As a consequence, a fused video combining the motion information from these two videos can provide a better way of estimating and interpreting the motion information in a scene.

In summary, the main contributions in this paper consist of the following:

1. We present an optical flow combination technique based on the certainties of the multi-source estimates.
2. We present a correlation-based motion selection procedure.

2 Optical Flow Technique

Because of its efficiency and accuracy, the local gradient method [1] is used here to extract optical flow. This technique relies on the assumption that the intensity values of the same pattern will be spatio-temporally conserved:

\[ I(x + \delta x, y + \delta y, t + 1) = I(x, y, t) \] (1)

where \( I(x, y, t) \) denotes the intensity value at the spatio-temporal coordinate \((x, y, t)\), \( \delta x \) and \( \delta y \) are the small displacements in \( x \) and \( y \) directions. By taking the Taylor expansion to \( I(x + \delta x, y + \delta y, t + 1) \) around the position \((x, y, t)\) and ignoring the influence from the higher order derivatives of the intensity function, the commonly used Optical Flow Constraint can then be written as:

\[ I_x u + I_y v + I_t = 0 \] (2)

where \( I_x \) and \( I_y \) are the spatial derivatives of the intensity function \( I \) at the position \((x, y, t)\) and \( I_t \) is its corresponding temporal derivative; \( u \) and \( v \) are the orthogonal components of the velocity vector in the \( x \) and \( y \) directions at the same position. In order to estimate the velocity vector, we therefore need to extract the spatio-temporal derivatives of image brightness. In measuring those derivatives a standard method is to convolve an image sequence with finite kernels spatio-temporally, and this procedure may introduce noise to the measurements. In order to obtain an analytical solution, we assume that only measurement of the temporal derivative \( I_t \) introduces noise \( \epsilon \) and that the noise is Gaussian distributed with zero mean and standard derivation \( \sigma \). Under such an assumption, we can construct the following stochastic model:

\[ \tilde{I}_t = I_x u + I_y v + \epsilon = a^T v + \epsilon \] (3)

where \( a = [I_x, I_y]^T \) and \( v = [u, v]^T \).

To obtain an optimal solution and avoid the aperture problem [6], neighbourhood information is usually considered. By further assuming that velocities for all pixels in a local region \( \Omega \) with \( N \times N \) size have the same velocity, a least squares estimation scheme is adopted to minimize:

\[ f(v) = \int_{x \in \Omega} (a^T v + \tilde{I}_t)^2 dx, \] (4)

and the final unbiased estimate of the vector \( v \) is given as:

\[ \bar{v} = -A^{-1}b \] (5)

where \( A = \sum (aa^T) \),

\[ b = \sum (\tilde{I}_t a) \].

Once a unique velocity estimate is obtained, we assign it to the central pixel of the local region. Thus, in order to estimate velocities for all pixels in an image, the standard procedure is to shift the central position of the region pixel by pixel throughout the whole image.

The above framework provides a simple way to recover the motion information in a video. However, least squares estimation is sensitive to outliers. Although modified methods, such as [7] and [8], have been proposed, they are computationally expensive. Instead, we adopt the following motion segmentation method to alleviate the influence of outliers. The apparent outliers in a local region occur mainly in two situations: (a) the co-existence of a single motion pattern and the background; (b) the co-existence of multiple motions. Regarding the former, as the temporal derivatives of the intensity function illustrate the change of illuminates over time, we can set a threshold to the derivative image so that only those positions whose absolute temporal derivatives are higher than the threshold are viewed as moving. Regarding the latter, we perform a local averaging procedure to the computed optical flow in the image so that the resultant velocity vectors in the motion boundary are either very small or similar to the leading motion in the region. Our IR camera (in common with many others) produces a black halo around hot objects. Whilst this helps to produce good segmentation of hot objects from their background, it introduces unwanted motion estimates around hot objects. In order to reduce the influence of those unwanted motion estimates, we set another threshold to the intensities of the IR videos so that only at those positions whose corresponding intensities in the IR video are higher than the threshold, the motion from the IR and the VIS videos will be further estimated.
In visualising the resultant optical flow estimates, we use grey-scale images to show the amplitudes of the velocities and colour images to illustrate the motion direction so that different orientations can be identified according to the specific colour in the colour wheel (see Section 4.1). A direct advantage of this method, compared to the traditional quiver plot, is that it allows a dense representation of local image velocity.

3 Multi-Source Videos

3.1 Video Data

In various applications, such as surveillance systems, patterns in the scene can be complex and the imaging conditions can be suboptimal. In order to test and develop algorithms that can cope with difficult or extreme conditions, a data gathering exercise was carried out by the University of Bristol and other partners at the Royal Fort Gardens in Bristol, UK, and at the Eden Project in Cornwall, UK. All the visible and infrared videos used in our work are from these two data sets. The multi-sensor system we used to capture the videos consists of a Ratheon Thermal eye 250D infrared camera and a Panasonic CCD AW-E300A visible camera, both of which were set to record at 25 frames per second. In addition to the example shown in Figure 1, some selected pairs of images from these videos are presented in Figure 2.

![Figure 2: Video examples from Royal Fort Gardens (top row) and the Eden Project (middle and bottom rows) data sets. Left column: infrared images, Right column: visible images.](image-url)

These videos illustrate typical scenes that challenge the detection and estimation of motion information from the visible camera because of influences from shadows (top row), complex surroundings (middle row) and dark imaging environments (bottom row). Our goal in this paper is to create an accurate and reliable representation of the motion information within these sequences.

3.2 Video Registration

Because of temporal and spatial misalignments, the raw videos from the multi-sensor system cannot be used directly and a spatio-temporal registration procedure is required. Regarding the temporal misalignment problem, whilst recording the videos we used a flash (which can be clearly detected by both sensors) as the starting signal. The occurrence of the spatial misalignment is largely due to the different physical properties of the sensors and the fact that although roughly collocated they are not at exactly the same position. The resultant pair of images can have different fields of view and optical distortions (Figure 2) and, usually, a geometric transformation is needed to map the frames of one of the videos to the other.

Because of its speed and relatively good performance, we use an affine transformation to align the videos. The affine transform parameters can be reliably estimated through a least squares estimation using a set of corresponding key points. As the video data are produced by a static multi-sensor system with fixed cameras, we can assume that the local transformations between these two sensors are constant over the recording time. Thus, key points used for the registration transformation can be chosen from different frame pairs. Although this assumption is generally valid, it works well only in cases where the distance in the scene between the moving objects and the sensors does not change much, otherwise we have a parallax problem and to solve it we need to update the registration transform parameters dynamically over the recording time [9]. Based on this idea, we have created a function in our real-time Video Fusion Toolbox (VFT) to allow users to add and delete pairs of key points from different frames. The VFT uses OpenCV and Intel’s Integrated Performance Primitives (IPP) and allows for real-time feedback so when a new key point is selected it immediately recomputes the affine transform coefficients and resamples the images so that the contribution each key point makes to the registration process can be seen. In Figure 3, we show the screenshots taken during the registration procedure. The subplots in (a) and (b) show a pair of unregistered images. The green points indicate the positions of the selected key points at current and past frames. The registered VIS image and the checkerboard representation are shown in Figure 3 (c) and (d) respectively. In the future we plan to implement higher order polynomial transforms for registration which will allow better compensation for lens distortions and reduce the errors due to parallax.
4 Optical Flow Based Motion Fusion

4.1 Optical Flow Combination

With our co-registered IR and VIS sequences, all the possible motion detected by the multi-sensor system can be classified into three categories:

1. Motion that can only be detected by the IR camera: motion of hot objects in the shadows and the darkness. All positions in this category are encoded as white colour in Figure 4 (a).

2. Motion that can only be detected by the VIS camera: motion of cold objects in daylight (Figure 4 (b))

3. Motion that can be detected by both the IR camera and the VIS camera. (Figure 4 (c))

In this section, we address the problem of combining motion information (optical flow) from the multi-source videos in order to provide a representation of all the motion detected in the scene. Regarding those motions which can only be detected by the IR or the VIS camera, their motion information will be directly integrated into the final representation. For motion information shared by the two cameras, we introduce the following combining method based on the certainties of the estimates.

Reviewing the optical flow technique presented in the Section 2, under the assumption that noise $e$ in the measurements of the temporal derivatives follow a Normal distribution $e \sim \mathcal{N}(0, \sigma^2)$, the final estimate in a local region is unbiased to the true motion vector $\mathbf{v} = [u, v]^T$:

$$E(\mathbf{v}) = E([\hat{u}, \hat{v}]^T) = \mathbf{v}$$  \hspace{1cm} (6)

where $E(\cdot)$ denotes the expectation. Correspondingly, the associated covariance matrix of the vector $\mathbf{v}$ has the following form:

$$\text{Cov}(\mathbf{v}) = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}$$  \hspace{1cm} (7)

where $\sigma_u^2$ and the $\sigma_v^2$ are the variances of $\hat{u}$ and $\hat{v}$ respectively, which measure the closeness or certainties of the components of the estimate comparing to their expectations. $\sigma_{uv}^2$ denotes the covariance between $\hat{u}$ and $\hat{v}$. Mathematically, the above covariance matrix can be computed according to [10]:

$$\text{Cov}(\hat{\mathbf{v}}) = \sigma^2 \mathbf{A}^{-1}$$  \hspace{1cm} (8)

where $\mathbf{A}$ is defined in Equation (5). In the above expression, as $\sigma^2$ is unknown, its unbiased estimate is required and is computed as follows.

Consider Equations (2) and (3), because $[\hat{u}, \hat{v}]^T$ is the optimal estimate of $[u, v]^T$, we have:

$$-\hat{I}_t = I_x \hat{u} + I_y \hat{v}$$  \hspace{1cm} (9)

where $\hat{I}_t$ is the estimate of the temporal derivative $I_t$. Based on this estimate, we can compute the estimate of the noise in the measurement of the temporal derivative according to:

$$\hat{e} = (-\hat{I}_t) - (-\hat{I}_t) = (-\hat{I}_t) - (I_x \hat{u} + I_y \hat{v})$$  \hspace{1cm} (10)
By taking all valid pixels in the region $\Omega$ into consideration, we can compute the following Sum of Squared Residuals (SSR):

$$SSR = \sum_{i=1}^{n} (\hat{e}^i)^2$$

(11)

where $n$ ($2 < n \leq N \times N$) is the number of pixels in $\Omega$ with motion patterns. Correspondingly, the estimate of $\sigma^2$ is computed according to:

$$\hat{\sigma}^2 = \frac{SSR}{n - p}$$

(12)

where $p$ is the number of unknowns in the regression and $p = 2$ in our case. By combining Equation (8) and (12), the estimated covariance matrix of an unknown vector is finally computed according to:

$$\widehat{\text{Cov}}(\hat{v}) = \begin{bmatrix} \hat{\sigma}^2_{u1} & \hat{\sigma}^2_{u2} \\ \hat{\sigma}^2_{v1} & \hat{\sigma}^2_{v2} \end{bmatrix} = \frac{SSR}{n - p} \mathbf{A}^{-1}$$

(13)

Similarly, if motion at a single location can be detected from both the IR and the VIS cameras, the individual estimates are all assumed to be the unbiased estimates of the ground-truth vector according to the least squares scheme:

$$E(\hat{v}_1) = v, \quad E(\hat{v}_2) = v$$

(14)

and correspondingly:

$$\widehat{\text{Cov}}(\hat{v}_1) = \begin{bmatrix} \hat{\sigma}^2_{u1} & \hat{\sigma}^2_{u2} \\ \hat{\sigma}^2_{v1} & \hat{\sigma}^2_{v2} \end{bmatrix}, \quad \widehat{\text{Cov}}(\hat{v}_2) = \begin{bmatrix} \hat{\sigma}^2_{u1} & \hat{\sigma}^2_{u2} \\ \hat{\sigma}^2_{v1} & \hat{\sigma}^2_{v2} \end{bmatrix}.$$  

(15)

At this stage, we therefore obtain two estimates of a true motion vector with different certainties. To obtain a unique representation of the true motion vector, the simplest method is to average these two estimates. However, without considering estimation certainties, an erroneous estimate could compromise the accuracy of the averaged velocity. To address this problem, we introduce the following combination rule.

In this rule, we aim to produce an unbiased estimate of the true motion vector in which the variances of velocity components are smaller than variances from either of the individual estimates. To this end, we consider velocity components in the $x$ direction and the $y$ direction separately and thus their correlations can be ignored. In this rule, we compute an optimal estimate of the velocity component in the $x$ direction as:

$$\hat{u}_o = \frac{\hat{\sigma}^2_{u1}}{\hat{\sigma}^2_{u1} + \hat{\sigma}^2_{u2}} \hat{u}_1 + \frac{\hat{\sigma}^2_{u2}}{\hat{\sigma}^2_{u1} + \hat{\sigma}^2_{u2}} \hat{u}_2.$$  

(16)

where $\hat{u}_o$ is the optimal estimate of $u$; $\hat{u}_1$ and $\hat{u}_2$ are two raw estimates from the multi-source videos. Similarly, in the $y$ direction, we have:

$$\hat{v}_o = \frac{\hat{\sigma}^2_{v1}}{\hat{\sigma}^2_{v1} + \hat{\sigma}^2_{v2}} \hat{v}_1 + \frac{\hat{\sigma}^2_{v2}}{\hat{\sigma}^2_{v1} + \hat{\sigma}^2_{v2}} \hat{v}_2.$$  

(17)

According to Equations (16) and (17), it can be easily shown that $\hat{u}_o$ and $\hat{v}_o$ are the unbiased estimates of $u$ and $v$. By using estimation certainties to compute the weights, an estimate with larger variance will be endowed with a smaller weight so that its influence on the combined velocity is decreased.

For $\hat{u}_o$ and $\hat{v}_o$, the unbiased estimates of their corresponding variances in the $x$ direction ($\hat{\sigma}^2_{u_o}$) and the $y$ direction ($\hat{\sigma}^2_{v_o}$) can be computed as:

$$\hat{\sigma}^2_{u_o} = \frac{\hat{\sigma}^2_{u1} \hat{\sigma}^2_{u2}}{\hat{\sigma}^2_{u1} + \hat{\sigma}^2_{u2}}, \quad \hat{\sigma}^2_{v_o} = \frac{\hat{\sigma}^2_{v1} \hat{\sigma}^2_{v2}}{\hat{\sigma}^2_{v1} + \hat{\sigma}^2_{v2}}.$$  

(18)

It can be seen above, that both the optimal variances in the $x$ direction and the $y$ direction are smaller than the variances of the individual estimates. The optimal estimate has higher estimation certainty than both individual estimates. In the above expression, we assume that data from the IR video and the VIS video are uncorrelated. In reality, the data may range anywhere from uncorrelated to fully correlated. In the latter, the combination procedure presented in Equations (16) and (17) is equivalent to a simple averaging procedure, and the variances of the individual estimates and the optimal estimate are identical. As the degree of correlation decreases the optimal estimate will increasingly have a higher estimation certainty than the individual estimates. In other words, unless the input sources are identical, the optimal estimate is always better than the individual estimates. How much better depends on the degree of correlation.

To examine performance, we perform a numerical comparison using some artificial sequences in which the ground-truth vectors are available [11]. To simulate the multi-source outputs, a low-frequency (LF) video is generated from the testing sequence by convolving each frame with a 2D Gaussian kernel $\kappa$, where $\rho$ is the standard deviation. Meanwhile, a high-frequency (HF) video is generated by subtracting the low-passed video from the testing video. In Figure 5 (a) and (b), a pair of videos generated from the Yosemite sequence is shown. In the experiment, $\rho$ varies from 1 to 5 and the window size of the neighborhood in which a constant velocity is assumed changes from $3 \times 3$ to $21 \times 21$ pixels. Mean Angular Errors (MAE) [11] (for a window size $11 \times 11$ pixels and $\rho = 3$) are shown in Figure 5 (c) and (d) respectively. These results clearly show that in most cases, the proposed procedure provides better performance than either of the two individual estimates. Although in some extreme cases, such as $\rho = 5$, the combination procedure may be less accurate than one of the single-source estimates, the combined estimate will always be close to the single-source estimate with the higher accuracy. This result is important as the combination procedure reduces the influence of bad estimates from a single video on the overall performance. In testing the performance using the Diverging Tree and Translating Tree sequences, the same conclusion can be drawn. In the future, we plan to perform such
numerical comparison using simulated IR and VIS images of artificial scenes where ground-truth data will be available.

(a) Low frequency video \( \rho = 3 \)
(b) High frequency video \( \rho = 3 \)

(c) MAE (window size is 11 × 11)
(d) MAE (\( \rho = 3 \))

Figure 5: Numerical comparison using the Yosemite sequence.

In Figure 6, the fused results from the multi-source videos are shown. The subplots in the first and the second rows show the estimated optical flow from the IR and the VIS videos individually. The representation of motion information using our combination method are shown in the third row. Additionally, the subplots in the bottom row show the certainties of the estimated shared by two videos in the \( x \) and \( y \) directions. Here, red indicates that the estimate from the IR video has higher certainty. In contrast, green indicates that the VIS estimate is more reliable. Some additional results for other video pairs are shown in Figure 7.

4.2 Motion Selection

The multi-sensor motion fusion technique presented in the previous section provides an accurate way to extract motion information from complex scene. If one wishes, additional analysis can then be performed on the extracted motion information. In many surveillance or military applications, interest lies in the motion information that characterises specific behaviours of moving objects. In the real world these motion characteristics will have some associated variability. It may well therefore be more meaningful to define a target in a fuzzy way rather than a deterministic way. To this end, we perform a correlation process by taking both the target velocity and its variation into consideration.

In this section, we propose a novel optical flow based motion selection procedure using a correlation technique. Rather than comparing an estimated velocity to the target velocity directly, the idea is to investigate the influence of the target velocity on a raw estimate through a combination procedure.
It is expected that a raw velocity estimate which is closer to the target velocity will be less influenced by the combination procedure so that this combination procedure introduces smaller changes to the raw estimate. In other words, this raw estimate will have higher similarity to the resultant combined vector incorporating the target velocity information.

In order to combine the target velocity information with the raw estimate, let us firstly review Equations (16) and (17). These two equations show that if we have a single estimate and a target velocity, and their corresponding certainties, we can construct a new vector based on this information. In Figure 8, we show an example of such a combination procedure. In this figure, \( \mathbf{v}_1 \) (red) and \( \mathbf{v}_2 \) (blue) are two estimates and \( \mathbf{v}_t \) (green) is the target velocity. By choosing a suitable variance for the target velocity, the combined velocity vectors \( \mathbf{v}_{1t} \) and \( \mathbf{v}_{2t} \) are the weighted summation of the raw estimates and the target vector. In the case that the raw estimate is close to the target, the weighted summation procedure introduces small difference between the new vector \( \mathbf{v}_{2t} \) and the raw estimate \( \mathbf{v}_2 \). In contrast, this combination will introduce large difference between the vector \( \mathbf{v}_1 \) and \( \mathbf{v}_{1t} \) as \( \mathbf{v}_1 \) is quite different from the target velocity. As a consequence, according to the similarity between the combined vector and the raw estimate, we can indirectly generate the closeness information between a raw estimate and the target velocity. In the case where the motion in a given direction is what we are interested in, we can simply define a series of target velocities with the same orientations, but different amplitudes. All the selected motion information from those target velocities can then be combined or overlapped to represent the motion information in a certain direction.

To measure the similarity, we compute the correlation between the components in the combined vector and the raw vector. The correlation in the \( x \) (\( C_x \)) and \( y \) (\( C_y \)) directions are computed as follows [9]:

\[
C_x = \frac{\sum_{\Psi} u_{1t} u_1}{\sqrt{\sum_{\Psi} u_{1t}^2 \sum_{\Psi} u_1^2}},
\]

\[
C_y = \frac{\sum_{\Psi} v_{1t} v_1}{\sqrt{\sum_{\Psi} v_{1t}^2 \sum_{\Psi} v_1^2}},
\]

where \( \Psi \) is the local region used to compute the correlation. Here \( C_x \) and \( C_y \) change from -1 to 1 and the closer to 1 they are, the higher the similarity is. Accordingly, a single estimate with high similarities in both \( x \) and \( y \) directions will be selected. In Figure 9, we show an example using the Diverging Tree sequence [11]. Here the velocity we are interested in is \( \mathbf{v}_t = [1, 1]^T \) and the variances in both \( x \) and \( y \) directions are set to \( 1e-5 \). In the top row of Figure 9,
above procedure can be applied to the fused optical flow produced from the technique introduced in Section 4.1. The only modification in this multi-source case is that the variances computed from Equation (18) will be used at those locations where motion can be detected by both sensors. In Figure 10, we show the results from this selection procedure. Here, the target velocities are set to $[-3, 0]^T$ and $[3, 0]^T$ respectively. The variance of the target velocity is set to one tenth of the variance of the estimated velocity vector. This setting allows more influence from the target velocity. The size of the region to compute the local correlation is set to $5 \times 5$ and the certainty is set to 0.99. Comparing to the optical flow representation shown in Figure 6 (f), it can be seen that the motion selection procedure filters the unwanted motions, leaving only motion patterns which are close to the target velocity. Meanwhile, the local correlation technique also helps to reduce the estimates of the random motion patterns in the leaves so that a clear representation is obtained.

![Figure 10: Motion selection results.](image)

(a) Moving leftwards  (b) Moving rightwards

5 Conclusion

We have presented a novel approach to extract the motion information from a multi-sensor (visible and IR) system. By combining the different advantages of our two kinds of sensor, the ability to detect and estimate local image motion can be enhanced. In dealing with those patterns which can be detected by both of the sensors, we introduced a probabilistic combination method based on the certainties of the estimates from the two sources. We have also introduced a novel correlation-based single source and multi-source motion selection procedure. This procedure allows us to incorporate application-specific prior knowledge and specify given motion requirements in a stochastic way. This in turn allows us to produce fused images in which motion close to a given target velocity is preferentially displayed. The advantage of our method is that it takes into account not only the difference between the estimates and the target, but also the certainty of the estimates. In our future research, the estimation framework will be extended to a multi-resolution structure in which large displacements can be accurately estimated and subsequently fused.

Acknowledgements

This work has been funded by the UK MOD Data and Information Fusion Defence Technology Center. The authors would like to thank John Lewis for his help with the initial video processing.

References


