Machine Vision and Applications (2000) 11: 267–276

Machine Vision and Applications
© Springer-Verlag 2000

MODEEP: a motion-based object detection and pose estimation method for airborne FLIR sequences

Alexander Strehl, J.K. Aggarwal*

Computer and Vision Research Center, The University of Texas at Austin, Department of Electrical and Computer Engineering, Austin, TX 78712-1084, USA; e-mail: {strehl,aggarwaljk}@mail.utexas.edu

Abstract. In this paper, we present a method called MODEEP (Motion-based Object DEtection and Estimation of Pose) to detect independently moving objects (IMOs) in forward-looking infrared (FLIR) image sequences taken from an airborne, moving platform. Ego-motion effects are removed through a robust multi-scale affine image registration process. Thereafter, areas with residual motion indicate potential object activity. These areas are detected, refined and selected using a Bayesian classifier. The resulting regions are clustered into pairs such that each pair represents one object's front and rear end. Using motion and scene knowledge, we estimate object pose and establish a region of interest (ROI) for each pair. Edge elements within each ROI are used to segment the convex cover containing the IMO. We show detailed results on real, complex, cluttered and noisy sequences. Moreover, we outline the integration of our fast and robust system into a comprehensive automatic target recognition (ATR) and action classification system.

Key words: Motion detection – Object segmentation – Pose estimation – Moving camera – Affine image registration – Infrared – Bayes

1 Introduction

1.1 Motivation

Forward-looking infrared (FLIR) images are frequently used in automatic target recognition (ATR) applications. ATR covers a variety of semi-automated and automated operations ranging from cuing a human observer to potential targets to fire-and-forget. Many researchers have investigated various approaches to detection, recognition and pose estimation of targets from *static* FLIR images. A comprehensive recent review by Ratches et al. on techniques for image-based ATR systems can be found in [RWBG97]. A variety of techniques to *detect* targets in static images have been proposed. Early work often was data-driven and used ad hoc methods such as thresholding based on the contrast of an object compared to the local background or pixel statistics. Later algorithms used knowledge-based systems and template-matching approaches. More recent research focuses on model-based approaches and multi-sensor fusion [NA92, RCM+95, BDZ+97]. While common ATR systems can track objects based on a series of single-frame detections, motion has been neglected as a cue to target detection and pose estimation. Motion information can be a very strong aid for finding objects in images as many biological vision systems indicate [Wan95]. So, rather than obtaining motion as a post-processing result of single-frame detection, we propose MODEEP, a method for motion-based object detection and estimation of pose. Including *dynamic* scene information in a static ATR system adds an independent criterion that can significantly increase detection rates and decrease false alarms.

Today, many techniques exist for the motion analysis of *visible-spectrum* imagery [BA96, AN89, MDD+95, IA98, TZ98]. Irani and Anandan differentiate scenes and the appropriate algorithms along a 2D to 3D continuum [IA98]. In 2D analysis, the scene can be approximated by a flat surface and the camera undergoes mainly rotations and zooms. 3D scenes are characterized by significant depth variations in the scene and a translating camera. Successful motion analysis requires using the appropriate model for the processing environment. That is, recovering structure from motion fails when features are sparse or the camera does not undergo sufficient translation.

In this paper, we present a motion-based object detection system tailored for FLIR sequences. Our FLIR sequences are taken from a moving platform and depict scenery as well as independently moving objects (IMOs). This case represents the most general scenario of motion because observer motion *and* object motions induce *multiple* coupled motions. In our approach, we compensate for the observer motion (ego-motion) that makes the background stationary. After removing the effects of ego-motion, any residual motions must be

* This research was supported in part by the Army Research Office under contracts DAAH04-95-I-0494 and DAAG55-98-1-0230 and by the Texas Higher Education Coordinating Board Advanced Research Project 97-ARP-275.
Correspondence to: J.K. Aggarwal

due to moving objects. We use these residual motion areas to detect and segment the objects and estimate their pose.

1.2 FLIR versus visible

To detect IMOs in FLIR image sequences, the sensor properties have to be taken into account. We face additional challenges caused by the following important differences to visual sequences.

- In FLIR imagery, an object's edges and corners appear smoothed out, reducing the number of distinct features in the image.
- The generation and the maintenance of kinetic energy usually heat up a moving object (e.g., friction, engine combustion). Consequently, moving objects often appear brighter than their environment in FLIR images.
- FLIR images are noisy and have less contrast. Moreover, they often contain artifacts such as dirt on the lens, brightness which fades out at the end of the scan line, or local sensor failure at certain pixel locations.
- FLIR sequences are not easily available (especially not from controlled experiments) and have a lower resolution. The sequences available to us are 128×128 pixels as compared to 512×512 pixels and more of standard visual cameras.
- FLIR sequences are often taken under difficult circumstances which result in abrupt discontinuities of motion.

These properties must be taken into account when building a successful system. The sparsity of distinct features and the noisiness of the data demand more robust techniques than are currently used for visual sequences.

1.3 Organization

Figure 1 shows a graphical overview of our proposed system. In the first module, we enhance the image quality to overcome problems such as low contrast, artifacts and noise (Sect. 2). Thereafter, we perform robust multi-scale affine image registration to eliminate effects from the motion of the camera platform (Sect. 3). Then, candidate regions for object parts are obtained by analyzing the residual misalignment. Using properties of the scene and the sensor, we remove unlikely regions and identify region pairs that correspond to the front and rear parts of the object. Together with edge elements, we obtain a convex cover for the IMOs' locations in the image (Sect. 4). Section 5 demonstrates experimental results and Sect. 6 summarizes the proposed system and suggests integration into a comprehensive ATR framework.

2 Image enhancement

Because FLIR images are inherently noisy and have less contrast, we first enhance image quality to facilitate further processing. In images recorded from a moving platform, artifacts appear as candidates for IMOs since they do not move coherently with the scene. To prevent this from leading to false alarms, the incoming frames are filtered before further

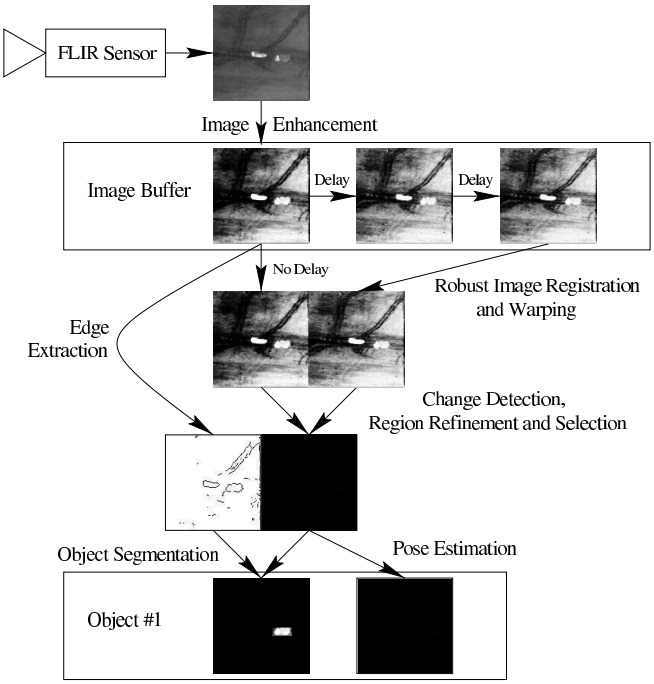


Fig. 1. Overview of MODEEP system

processing. Locations with artifacts often have completely erroneous gray-level values. In the case of salt-and-pepper noise, order statistic filters provide a model-free solution, more suitable than non-robust filters. We use a median filter to successfully remove small artifacts and image noise while preserving relevant edge information.

FLIR images are based on the thermal electro-magnetic spectrum. Differences within a scene's background are rather small compared to differences between background and objects. This leads, in general, to a very low contrast in most of the image area. The second enhancement, histogram equalization, normalizes contrast and compensates for various base brightness levels. Histogram equalization re-maps gray levels in an order-preserving fashion such that the cumulative histogram has an approximately linear slope. In the next section, we will discuss how the effects of camera motion are removed from the enhanced sequence.

3 Robust multi-scale affine registration

Moving objects induce motion in an image sequence. Since their image motion is different from the image motion caused by the camera's movement, they are referred to as IMOs. In our case of airborne imagery, the objects are moving on the ground and appear rather small. Consequently, the background of the scene will cover most of the image. The dominant motion is a motion that explains most of the apparent motion. The background in the image undergoes displacement caused by the observer's movement (or ego-motion) and, hence, constitutes the dominant motion. IMOs can also be understood as objects whose motion violates the dominant motion model. In order to detect such objects, we remove the effects of the prevailing (dominant) motion from the sequence. This leaves only the effects of secondary and smaller motions (the independent motions).

Due to high noise and frequent large displacements, we have to use the entire image and cannot rely on a windowed approach to compensate for motion. Using a 3D motion model to compensate requires a depth map from the scene. This depth map can be either given or estimated from the sequence, if sufficient scene texture and translational ego-motion are present [Adi85, IA98]. While 3D models have a small bias (expected model error), they are prone to high estimation error (variance) due to their high number of degrees of freedom (one unknown depth parameter for each location in the image plus rigid-motion parameters). Considering FLIR shortcomings and the noise sensitivity of motion estimation, the 2D affine model with its six degrees of freedom provides a good balance for the bias-variance tradeoff. An estimator is robust if outliers cannot arbitrarily worsen the estimate. By applying robust statistics [Hub81] to motion estimation [BA96], the dominant motion estimate can be made invariant to small model violations such as IMOs or minor depth discontinuities (parallax). The selection of the motion model is crucial to the success of compensating for camera motion.

There are several ways to estimate and compensate for the dominant observer motion. Feature-based motion estimation [Wu95, BB90, TZ98] seems inappropriate because very few features are present. Also, these tend to be IMOs and would thus skew the ego-motion estimates. Abrupt discontinuities in the motion as a result of camera movement make spatio-temporal filtering approaches [WA94] ineffective, too. The best method appears to be a registration technique that uses the entire image and is able to handle large displacements while being robust against the violations in object motion. Since the moving objects are very small in airborne images (maximally 10% of the image area), we can assume that camera motion is the dominant motion in the scene.

Let $\mathbf{I}^t$ represent the image intensity as a function of the image location $\mathbf{x} = (x_1, x_2)^\mathbf{T}$ at time t and $\theta^{t-\tau}$ the motion parameter vector describing the visual motion from the frame at time $t-\tau$ to the next (at time t).

$$\mathbf{x}^t = \mathscr{M}^{t-\tau}(\mathbf{x}^{t-\tau}, \theta^{t-\tau}) \; . \tag{1}$$

For our system, we use the entire image in a robust multi-scale affine image registration [BAHH92]. This aligns a frame $\mathbf{I}^{t-\tau}$ to a reference frame $\mathbf{I}^t$, assuming an affine transformation of the homogeneous coordinates [FvDFH90] as described in Eq. 2. We always use the most recent frame available as the reference frame (in contrast to always using the first frame of a sequence).

$$\mathscr{M}(\mathbf{x}, \theta) = \begin{pmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} \; . \tag{2}$$

The 2D affine motion model has six parameters as seen in Eq. 3.

$$\theta = (\theta_1 \; \theta_2 \; \theta_3 \; \theta_4 \; \theta_5 \; \theta_6)^\mathbf{T} \; . \tag{3}$$

The motion transformation $\mathscr{M}$ is estimated in four stages [BAHH92], as described in the following subsections.

3.1 Pyramid construction

A Laplacian image resolution hierarchy is created to allow processing on various spatial frequency levels. In a Laplacian pyramid, the image is decomposed into one low-resolution low-pass-filtered image and multiple higher resolution layers encoding the higher frequencies [BA83]. We start motion estimation at the lowest resolution level of 32×32 and expand and refine the results layer by layer until the original resolution of 128×128 is reached.

3.2 Motion estimation

Most motion estimation paradigms are based on image intensity conservation. Intensity conservation assumes that during a sufficiently small time τ between frames, no intensity pattern in the image is lost. However, it may become displaced by u_1 and u_2 in x_1 and x_2 directions as expressed by Eq. 4, which was initially proposed by Horn and Schunk in [HS81].

$$\mathbf{I}^{t-\tau}(\mathbf{x}) = \mathbf{I}^t(\mathbf{x} + \mathbf{u}^{t-\tau}(\mathbf{x})) \; . \tag{4}$$

In each layer of the Laplacian pyramid, motion is estimated. We use an iterative estimator for θ that minimizes the sum of squared differences (SSD) between the reference frame $\mathbf{I}^t$ and the registered frame $\hat{\mathbf{I}}^{t-\tau} = \mathscr{M}(\mathbf{I}^{t-\tau}, \theta)$.

$$\hat{\theta} = \min_\theta(\mathrm{SSD}(\mathbf{I}^t, \mathscr{M}(\mathbf{I}^{t-\tau}, \theta))) \; . \tag{5}$$

The SSD is an error measure between two images $\mathbf{I}$ and $\mathbf{J}$ based on the intensity conservation assumption [HS81] and defined as follows:

$$\mathrm{SSD}(\mathbf{I}, \mathbf{J}) = \sum_\mathbf{x} (\mathbf{I}(\mathbf{x}) - \mathbf{J}(\mathbf{x}))^2 \; . \tag{6}$$

The initial (iteration $n = 0$) motion guess is 'no motion' and, hence, the motion model is identity $\mathscr{M}(\mathbf{x}, \hat{\theta}_0) = \mathbf{x}$, which is obtained with $\hat{\theta}_0 = (1\,0\,0\,0\,1\,0)^\mathbf{T}$. Using the Gauss-Newton method to minimize the SSD error with respect to the motion parameters θ, we obtain an incremental parameter update δ_n as given by Eqs. 7, 8, and 9:

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \delta_n \; , \tag{7}$$

$$\delta_n = -\left(\sum_\mathbf{x} \mathbf{P}^\mathbf{T} (\nabla \mathbf{I}^{t-\tau})(\nabla \mathbf{I}^{t-\tau})^\mathbf{T} \mathbf{P} \right)^{-1} \cdot \cdot \left(\sum_\mathbf{x} \mathbf{P}^\mathbf{T} (\nabla \mathbf{I}^{t-\tau})(\Delta \mathbf{I}_n) \right) \; , \tag{8}$$

$$\mathbf{P} = \begin{pmatrix} x_1 & x_2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_1 & x_2 & 1 \end{pmatrix} \; . \tag{9}$$

The residual error $\Delta \mathbf{I}_n$ is computed as the pixel-wise difference between the reference frame and the registered frame $\Delta \mathbf{I}_n = \mathbf{I}^t - \hat{\mathbf{I}}_n^{t-\tau}$. The image gradient $\nabla \mathbf{I}^{t-\tau}$ is approximated by filtering the image with the Sobel kernel [Sob70] for the horizontal and vertical direction (Fig. 2).

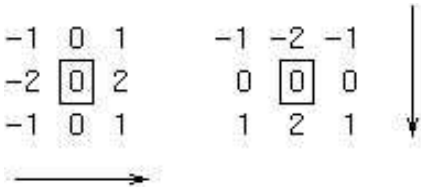


Fig. 2. Sobel edge filter. Linear filter kernels for x_1- and x_2-direction

3.3 Image warping

The current dominant motion estimate $\hat{\theta}_n$ at iteration n is used to warp the earlier image $\mathbf{I}^{t-\tau}$ to match the reference image $\mathbf{I}^t$. We employ a standard warping technique using bilinear interpolation. Bilinear interpolation defines the gray level at an intermediate locations (x_1, x_2) between actual pixel values (i, j) as the linear combination of its four nearest neighbor intensities weighted by their distance.

$$\begin{aligned}
\mathbf{I}(x_1, x_2) &= (1 - x_2 + j) \cdot \mathbf{I}(x_1, j) + \\
&\quad (x_2 - j) \cdot \mathbf{I}(x_1, j+1) \,, \\
\mathbf{I}(x_1, j) &= (1 - x_1 + i) \cdot \mathbf{I}(i, j) + \\
&\quad (x_1 - i) \cdot \mathbf{I}(i+1, j) \,, \\
\mathbf{I}(x_1, j+1) &= (1 - x_1 + i) \cdot \mathbf{I}(i, j+1) + \\
&\quad (x_1 - i) \cdot \mathbf{I}(i+1, j+1) \,. \qquad (10)
\end{aligned}$$

The warped image $\hat{\mathbf{I}}_{n+1}^{t-\tau} = \mathscr{M}(\mathbf{I}^{t-\tau}, \hat{\theta}_{n+1})$ is used instead of the original frame $\hat{\mathbf{I}}_n^{t-\tau}$ and the motion estimation process is repeated at iteration $n+1$. Motion estimation and image warping are iterated with the updated image $\hat{\mathbf{I}}_{n+1}^{t-\tau}$ and the reference frame $\mathbf{I}^t$. Iteration is terminated upon reaching a fixpoint for the motion estimate ($\delta_n \to \mathbf{0}$) or a maximum number of iterations. The selection of the maximum number of iterations depends on the expected magnitude of inter-frame motion (typically 3 – 10 iterations).

3.4 Refinement

The estimates are refined by expanding the results within the resolution pyramid in a coarse-to-fine fashion (Fig. 3). Since a Laplacian resolution hierarchy is used, image width and height double when stepping down a layer and, consequently, the motion estimation process at the higher resolution level is initialized with $2\hat{\theta}$. This prevents aliasing of high-spatial-frequency components that undergo large motions and minimizes outlier sensitivity. It also speeds up the motion analysis, since fewer iterations are required at each resolution level [BAHH92, BAK91].

4 Locating moving objects

4.1 Change detection and region refinement

After the effects of camera motion have been removed, the remaining regions with significant changes may contain IMOs. To determine which regions exhibit significant

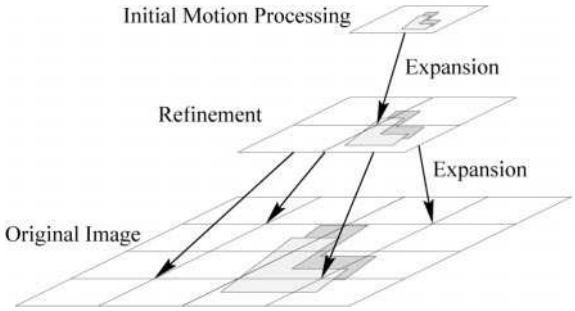


Fig. 3. Illustration of coarse-to-fine motion processing

change, we first compute the difference of the current image (the reference image) to a registered frame from the past (e.g., 0.2 s ago or 5 frames at 25 frames per second). The time difference between the frames must be long enough for the IMO to move significantly between them (e.g., its movement is detectable in the image, considering resolution and target distance). Locations exceeding a certain threshold in absolute difference are considered outliers to the background motion and constitute our initial change regions. Depending on whether the difference is significantly positive or negative, the initial change regions are elements of $\mathbf{I_h} = \text{bin}_+(\mathbf{I}^t - \mathscr{M}(\mathbf{I}^{t-\tau}, \theta^{t-\tau}))$ or $\mathbf{I_t} = \text{bin}_-(\mathbf{I}^t - \mathscr{M}(\mathbf{I}^{t-\tau}, \theta^{t-\tau}))$, respectively. These are then processed with morphological operations such as erosion (each pixel adopts the lowest value in its neighborhood) and dilation (each pixel adopts the highest value in its neighborhood). Iterative application of opening operations (erosion followed by dilation) in a 3×3 neighborhood smoothes the contours of regions, breaks narrow isthmuses and eliminates protrusions. The opening operations are repeated until the image no longer changes. A final dilatation operation grows the remaining regions.

$$\mathbf{I_H^t} = \text{dilate}(\lim_{n\to\infty}(\text{open}^n(\mathbf{I_h})) \,, \qquad (11)$$

$$\mathbf{I_T^t} = \text{dilate}(\lim_{n\to\infty}(\text{open}^n(\mathbf{I_t})) \,. \qquad (12)$$

The joint set of regions $\mathbf{I_r}$ of the two resulting images $\mathbf{I_H}$ and $\mathbf{I_T}$ contains the candidate regions for IMO parts.

4.2 Region selection and pose estimation

Some candidate regions may not correspond to a moving object. For example, heavy noise, artifacts or partial sensor failure could induce such false-alarm regions. To eliminate false alarms, we compute four features $s_{i,1}$ to $s_{i,4}$ for each candidate region in $\mathbf{I_r}$. The symbol X_i denotes the set of points in a particular region i. The r-th (central) momentum of X_i is denoted $m_r^{(0)}$ ($m_r^{(1)}$).

$$m_r^{(c)} = E\left[(X_i - c \cdot E[X_i])^r\right] \,, \qquad (13)$$

$$s_{i,1} = \; m_1^{(0)} \; = \text{mean} \,, \qquad (14)$$

$$s_{i,2} = \; m_2^{(1)} \; = \text{variance} \,, \qquad (15)$$

$$s_{i,3} = \frac{m_3^{(1)}}{\sqrt{m_2^{(1)}}^{\,3}} = \text{skewness} \,, \qquad (16)$$

$$s_{i,4} = \frac{m_4^{(1)}}{\sqrt{m_2^{(1)}}}^4 = \text{kurtosis} \; . \tag{17}$$

Based on these features, we decide if a region will be processed further or rejected as a false alarm. Due to the severely deteriorated image quality at the right and lower borders (end of scan line), we want to reject regions centered very close to any image margin. Moreover, size, symmetry and compactness can be used to exclude other false alarms. All of these properties are captured by the four region features. We use them in a Bayesian approach [DH73] to make a decision κ_i regarding the selection or rejection of a candidate region i based on its likelihood of being caused by a moving object.

The *a posteriori* probability that the region i is part of a target, given its feature vector $\mathbf{s}_i$, is denoted $P(T_1|\mathbf{s}_i)$. The *a posteriori* probabilities $P(T_k|\mathbf{s}_i)$ are computed using Bayes' rule and the law of total probability as shown in Eq. 18.

$$P(T_k|\mathbf{s}_i) = \frac{p(\mathbf{s}_i|T_k) \cdot P(T_k)}{\sum_h p(\mathbf{s}_i|T_h) \cdot P(T_h)} \; . \tag{18}$$

The probability densities $p(\mathbf{s}_i|T_k)$ are assumed to be multivariate (e.g., 4D) Gaussian densities:

$$p(\mathbf{s}_i|T_k) = \frac{\sqrt{|\Sigma_k^{-1}|}}{(2\pi)^2} \exp\left(-\frac{1}{2}(\mathbf{s}_i - \mu_k)^{\mathbf{T}} \Sigma_k^{-1} (\mathbf{s}_i - \mu_k)\right) \; . \tag{19}$$

Their parameters μ_k and Σ_k are computed as maximum-likelihood (ML) estimates [DH73] from dedicated training examples. Within a comprehensive ATR system, we plan to use a static detector to provide labeled examples for the training. The *a priori* probabilities $P(T_k)$ are also obtained from the training data as the relative frequencies of false alarms and targets.

$$\kappa_i = \begin{cases} \text{accept if} & \arg\max_k(P(T_k|\mathbf{s}_i)) = 1 \text{ and} \\ & P(T_1|\mathbf{s}_i) > \beta \\ \text{reject else} & \end{cases} \tag{20}$$

For each region, we decide if the region corresponds to an object (accept) or a false alarm (reject). The decision κ_i is made according to the decision rule (Eq. 20) based on the *a posteriori* probabilities and the confidence threshold β (e.g., 90%). If false alarms are to be avoided, β should be increased. Conversely, if missed detections have a high cost, β should be decreased. The appropriate value for β depends on the cost of a false alarm compared to a missed detection. All regions not meeting the minimum confidence requirement β are unlikely to be moving objects. Hence, these are rejected and removed for further processing. The remaining regions are the final IMO part regions. Through the growing process they include the adjacent boundaries of the corresponding objects.

We call the foremost part of the IMO in the direction of its movement the front part, and the opposite end its back. Since the sequences are recorded from airborne sensors, we are significantly above the plane on which the targets move. This assures that the front and back parts of regular vehicles cannot be hidden due to self-occlusion. One key property of infrared sensors is that targets or their parts (especially their *hot spots* such as the engine and the exhaust) appear brighter than the background. We can distinguish four cases of object motion and their resulting FLIR inter-frame intensity changes:

object is	in front of object	behind object
appearing	becomes brighter	not observable
moving visible	becomes brighter	becomes darker
disappearing	not observable	becomes darker
moving occluded	not observable	not observable

We call IMO regions heads (tails) if intensity increased (decreased) significantly from the registered to the reference frame. Heads are likely to contain an object's front, and tails usually indicate regions that an object's back has just vacated. Head and tail regions indicate the location of an object as well as its pose relative to the observer.

In case of multiple moving objects, we must find pairs of final regions corresponding to the *same* moving object. This also helps eliminate misdetected regions (false alarms), since it is very unlikely that there is a matching opposite region to form a valid pair. We cluster the detected regions into pairs consisting of a head and a tail region. To establish pairs, we assume that the distance from one object's front to its tail is smaller than from any of its parts to the contrary part of any another object. All possible pairs (combinations of a head X_i from $\mathbf{I_H}$ and a tail X_j from $\mathbf{I_T}$) are considered and ranked by the distance measure $p_{i,j}$:

$$p_{i,j} = |s_{i,1} - s_{j,1}| \; . \tag{21}$$

Starting from the closest match (lowest ranking), we now successively assign two regions to one pair. Since each region can be in only one pair, this accomplishes the desired clustering. Excess head or tail regions (false alarms) remain unpaired and are dropped at this stage. Equation 22 gives a more formal description of the clustering procedure:

$$\lambda_{i,j} = \begin{cases} \text{accept if} & p_{i,j} < p_{k,j} \; \forall k \neq i \text{ and} \\ & p_{i,j} < p_{i,l} \; \forall l \neq j \\ \text{reject else} & \end{cases} \; . \tag{22}$$

The decision $\lambda_{i,j}$ indicates if the pair formed by head region X_i and tail X_j is considered a valid final IMO pair.

Let us assume that a matching pair of a head and tail region has been found. We can approximate the object's pose in the image by the angle α' of the straight line from the centroid of the object's head $s_{i,1}$ to the centroid of the tail $s_{j,1}$. In our notation, $\alpha' = 0$ and $\alpha' = 90$ represent the directions straight up and straight to the right, respectively. For the typical airborne surveillance application, let us assume an elevated camera with a large focal distance ($f = \infty$ or parallel projection) looking forward at an object on a planar surface as depicted in Fig. 4. This scene geometry can be used to link the image pose α' to the true object pose α. The true pose α is defined here as the direction of the vehicle's heading on the ground plane in respect to the observer:

$$\tan(\alpha) = \tan(\alpha') \cdot \sin(\gamma) \; , \tag{23}$$

$$\text{distance} \cdot \tan(\gamma) = \text{altitude} \; . \tag{24}$$

These equations can be rewritten to obtain a universal closed-form solution for α using the function atan, which

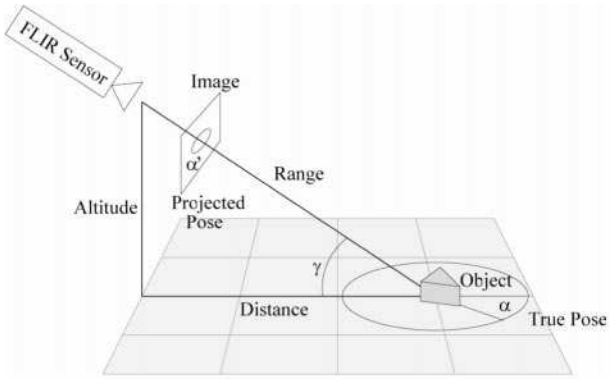


Fig. 4. Scene geometry for planar surface and elevated observer

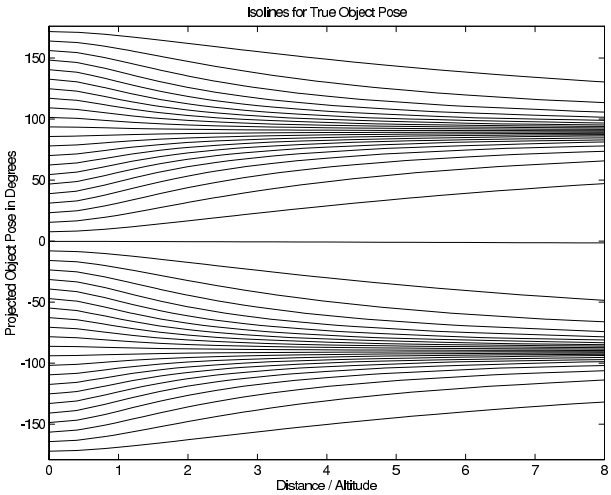


Fig. 5. Contour plot of the true object pose as a function of the distance-to-altitude ratio and the projected object pose α'. *Lines* show locations of equal true object pose α

is a generalized arctan function that computes an angle from a vertical and a horizontal component. Approximate knowledge of the camera's elevation above the ground plane (altitude) and its distance to the object on the ground allows us to compute α as follows:

$$\alpha = \text{atan}(\sin(\alpha') \cdot \sin(\text{atan}(\text{altitude}, \text{distance})), \cos(\alpha')) \; . \quad (25)$$

Figure 5 shows a contour plot of the true object pose as a function of the distance-to-altitude ratio and the projected object pose in degrees. The true pose angles $\alpha = 0$, $\alpha = 90$, $\alpha = 180$, $\alpha = -90$ correspond to the vehicle pointing outbound, to the right, inbound, and to the left with respect to the observer. For distance/altitude = 0, the observer is exactly above the object and, hence, perceives the true pose ($\alpha' = \alpha$). With increasing distance at constant altitude, the motion component in z-direction becomes less visible. In the limit, only strict left (−90 degrees) and right (90 degrees) movement can be perceived. This graph also shows that in high distance/altitude scenarios, small image pose estimation errors around 90 and −90 degrees result in large true pose estimation errors. From a long distance, it is hard to visually estimate if an object is moving in- or outbound.

4.3 Edge extraction and segmentation

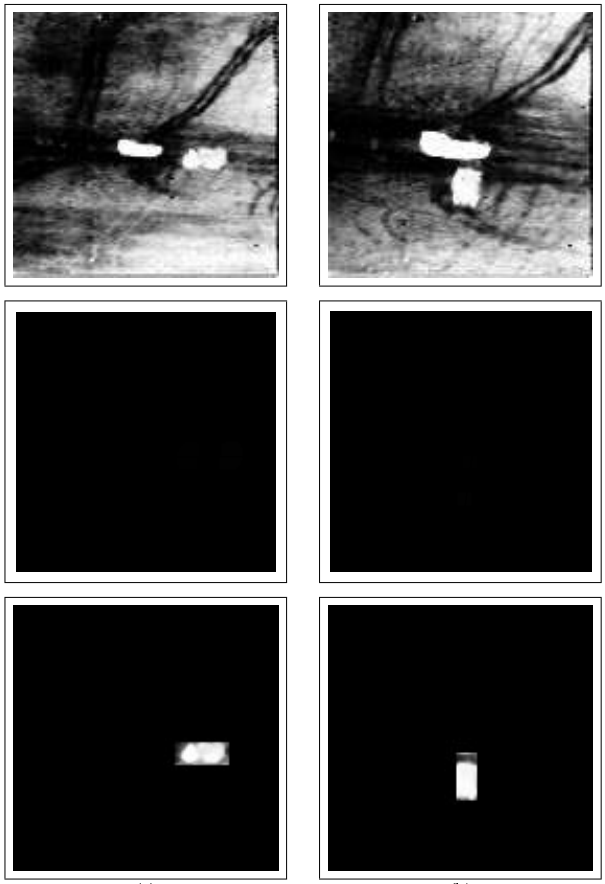
As we have just seen, the IMO regions indicate the front or rear part of the moving object. However, not all parts of the object are included into these two sets of regions. Motion of homogeneously intense areas, for example, cannot be observed. How can we find the entire object from the IMO part pairs? We have to resort to another feature domain, since pure motion information is not sufficient to solve this problem. Gray-level edges in the image can provide an indication of an object's boundaries. Independently from the motion detection, we extract the edges from the reference frame using the Canny operator [Can86] and the Sobel approximation of the derivative [Sob70]. This can be done in parallel with the change detection. At this point we assume that the objects project to convex regions in the image with (eventually only partially) visible object boundaries in the direction of motion. Even though the convexity assumption may not hold for all objects, its violation leads to the detection of the convex part, which is usually sufficient. Since the IMO regions were grown, they now include the object boundaries. The edge in the head region corresponds to the IMO's front end, and the edge in the tail region to the rear end. Consequently, locations fulfilling both constraints, lying within an IMO region and classified as an edge location, are the boundary locations of the object. Using the convexity assumption, the convex cover of the boundary regions constitutes the desired ROI containing the IMO.

5 Results

5.1 Tank-and-truck (TAT) sequence

Figure 6 shows two FLIR frames (top row) and detected IMOs (bottom row) and Fig. 7 depicts a spatio-temporal view of the entire sequence. During frames 1 to 30, a truck (the IMO) approaches the tank that sits in the center of the image. The elevated camera gradually comes closer. At frame 34, the camera was struck, resulting in an abrupt spatio-temporal discontinuity of the data. The camera fixates back at frame 38, but until the last frame 79, the sequence is unstable, with large inter-frame displacements. During this interval, the truck stops briefly and changes direction, driving toward the observer. The sequence demonstrates a mixture of continuous translational and abrupt, unsteady rotational camera motion.

Figure 8 illustrates the success of the frame-wise registration to stabilize the sequence. Various spatio-temporal slices through the entire sequence are shown before and after stabilization. In each slice the time progresses towards the right, and the upright axis is the free spatial axis. The stabilization removed the small and short-term effects of the wobbling camera (the jagged lines in Fig. 8a become smooth in b), as well as the continuous effect of the camera coming closer (the diverging lines in Fig. 8a become parallel in b). It is interesting to note the merging of the bright traces of the sitting tank and the moving truck in Fig. 8c and d. In Fig. 8a and b, the IMO 'enters' the vertical slice late in the sequence and appears as the lower chip of the bright trace.

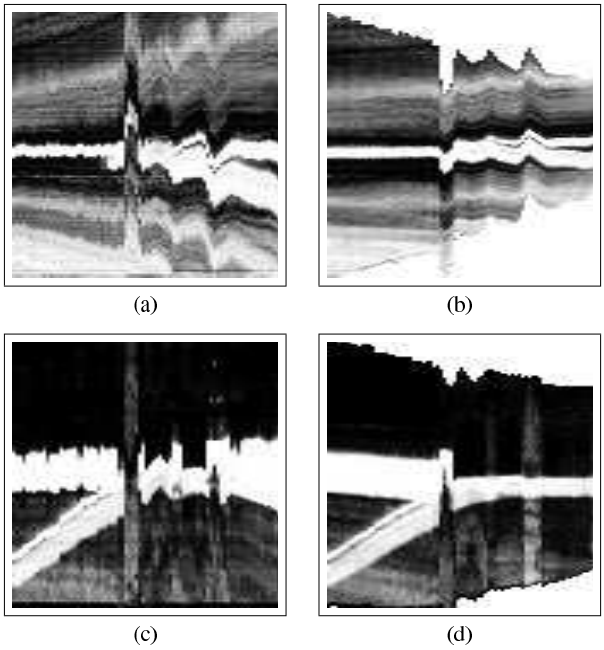


(a) (b)

Fig. 6a,b. Detection and pose estimation results obtained with our system. TAT FLIR frames 15 **a** and 72 **b** original (*top row*) and the final object part pairs with pose arrows (*middle row*). The *bottom row* shows the corresponding ROIs

Fig. 7. Spatio-temporal view of the TAT sequence. The data volume's slices at $t = 1$, $t = 79$, $x = 60$, and $y = 60$ are shown

The middle row of Fig. 6 shows the detected head (black) and regions behind the object's tail (gray). The objects' estimated direction of movement α' is indicated by the arrows. The final object segmentation obtained from edge and motion information is shown in the bottom row of Fig. 6. In frames 15 and 72, the IMO is located accurately and successfully segmented from the stationary components of the scene. While our system reliably detects the IMO for most frames in the sequence, it fails in frame 34 when the camera



(a) (b) (c) (d)

Fig. 8a–d. Effectiveness of ego-motion compensation. Vertical slices at $x = 60$ (*upper row*) and horizontal slices at $y = 60$ (*lower row*) before (*left column*) and after stabilization (*right column*) of TAT sequence

is struck heavily. This induces an abrupt and large displacement of the entire scene that cannot be compensated with the registration module. Consequently, many scene features appear as candidate parts and no objects are detected.

Figure 9 shows several intermediate processing results for frame 15. In Fig. 9a, the original pixel-wise difference of the current reference frame 15 and the previous frame 8 is shown. The difference depicted ranges from black (strong decrease) over gray (no change) to white (strong increase). After multi-scale registration, the observer motion is removed and the errors in the difference image (Fig. 9b) are due to IMOs. The initial regions for IMO parts (Fig. 9c) are refined through morphological operations to obtain the candidate IMO part regions (Fig. 9e). Candidate regions are selected (which in this case removes the false alarm regions at the margin) and paired. Edges (Fig. 9d) within valid pair regions constitute the IMO boundaries as shown in Fig. 9f. From an overall perspective we obtain excellent results, especially when considering the quality of the FLIR sequence. The vehicle is detected and segmented successfully and accurately during 47 of 72 frames of the TAT sequence (79 total at 7 frames corresponding to τ). Figure 10 shows the obtained pose estimates based *only* on image motion. The estimated pose changes from approximately -85 to -175 degrees. This correctly represents the trucks' left turning action.

5.2 Other illustrative examples

Results for three other complex and difficult sequences are shown in Fig. 11. The top row shows the reference FLIR frame. The difference to the previous frame before registration is depicted in the second row to illustrate overall motion effects. The third row illustrates the residual differences af-

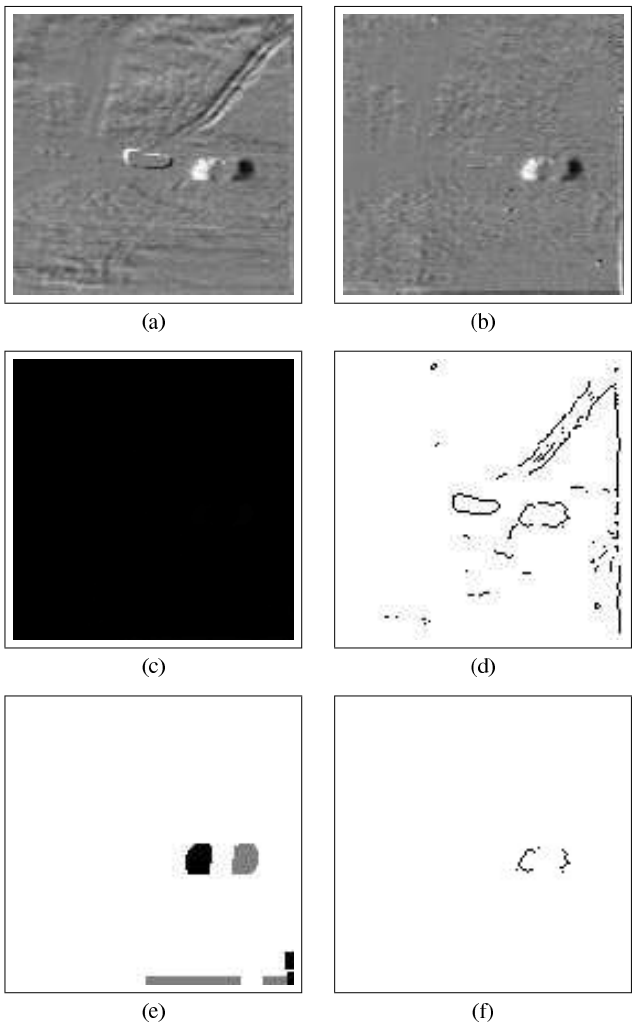


Fig. 9a–f. Steps in the processing at TAT frame 15. **a** Original difference of frames 8 and 15. **b** After affine multi-scale registration. **c** Initial IMO parts. **d** Edge map. **e** Candidate IMO parts. **f** Final IMO boundary parts

ter dominant motion compensation. In the bottom row, the candidate parts are shown. Final part pairs are overlaid with a pose indication arrow.

The left column of Fig. 11a shows a frame from a sequence containing two sitting tanks and no moving objects. Despite the large depth variations, motion compensation is successful and no false alarms are induced.

The sequence in the middle column (Fig. 11b) shows a tank moving across an unobstructed field towards the observer. The system successfully detects the heated right wheels and gives a good estimate of the tank pose. It can also be seen that the hot exhaust fumes induce a false alarm by appearing to be a head part. However, the fumes do not follow rigid motion. A heat edge appears on the fumes' front, but, due to the gradual dilution and cooling, no corresponding tail exists. Consequently, the falsely detected head remains unpaired and is rejected.

The frame shown on the right (Fig. 11c) contains two moving objects in a highly cluttered scene (road and trees), a tank moving rapidly to the right and another object moving towards the upper left of the image. Static ATR systems and even human observers may have difficulties detecting

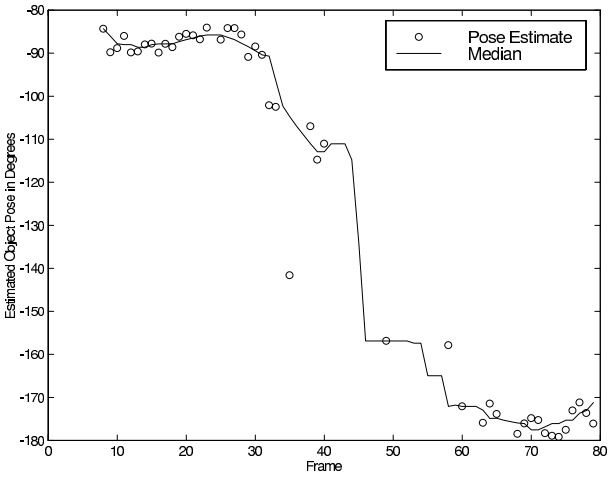


Fig. 10. Pose development during turning action of truck

the targets in this image. Our dynamic system successfully detects both objects' head and tail and recovers their poses. A false-alarm tail in the upper right corner is also detected at first, but is rejected later, since it cannot be paired.

5.3 Evaluation model

From the user's perspective, an object detection system has two modes of failure, namely false alarm and missed detection. To provide an objective basis for an evaluation and comparison of different object detection systems, we have to measure the true number of objects $|T|$, the number of detected objects $|D|$, and the number of correctly detected objects $|C| \leq \min(|T|, |D|)$ for each frame. Consequently, the number of missed detections is $|T| - |C|$ and the number of false alarms is $|D| - |C|$. A relative indication of the system's correctness for a single frame is then given by

$$\eta = \begin{cases} \frac{|C|}{|T|+|D|-|C|} & \text{if } |C| \neq 0 \text{ or } |D| \neq 0\,, \\ 1 & \text{else}\,. \end{cases} \tag{26}$$

Furthermore, we define the missed detection rate as

$$\varepsilon_1 = \begin{cases} \frac{|T|-|C|}{|T|+|D|-|C|} & \text{if } |C| \neq 0 \text{ or } |D| \neq 0\,, \\ 0 & \text{else}\,, \end{cases} \tag{27}$$

and the false alarm rate as

$$\varepsilon_2 = \begin{cases} \frac{|D|-|C|}{|T|+|D|-|C|} & \text{if } |C| \neq 0 \text{ or } |D| \neq 0\,, \\ 0 & \text{else}\,. \end{cases} \tag{28}$$

These are conservative performance estimators, with $0 \leq \eta, \varepsilon_1, \varepsilon_2 \leq 1$ and $\eta + \varepsilon_1 + \varepsilon_2 = 1$. We tested our framework on four real sequences with 1183 frames. Cumulative absolute and average relative performance metrics are given in Table 1. The expressiveness of the performance metrics is limited due to the strong variation in sequence quality and clutter. Please note that the given rates are per single frame. Assuming stochastic independence of frames, given a detection probability of 0.40, the target is detected in at least one of ten consecutive frames with probability 0.99. Considering that only motion information is used and targets often are not larger than ten pixels, we obtain excellent detection results.

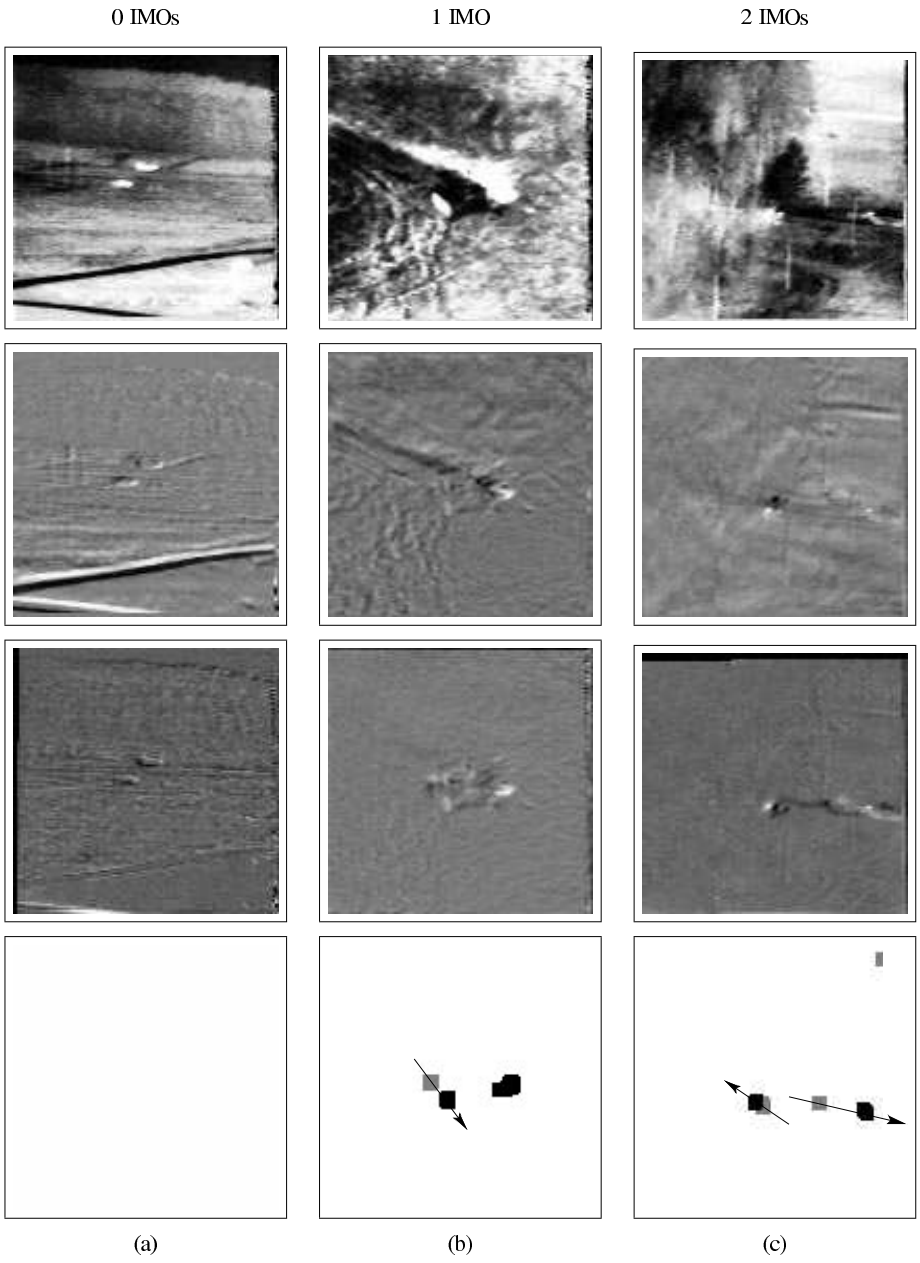


Fig. 11a–c. More results on complex and difficult sequences. Reference frames (*top row*), differences before (*second row*), after alignment (*third row*), and candidate parts with superimposed pose estimates for final pairs (*bottom row*)

Table 1. MODEEP performance metrics

targets	frames	$\Sigma\|T\|$	$\Sigma\|D\|$	$\Sigma\|C\|$	$\bar{\eta}$	$\bar{\varepsilon}_1$	$\bar{\varepsilon}_2$
0	165	0	25	0	92%	0%	8%
1	507	507	691	327	56%	37%	7%
2	511	1022	1231	711	49%	42%	9%
total	1183	1529	1947	1038	58%	34%	8%

6 Conclusion and future work

In this paper we propose MODEEP, a novel motion-based object detection system for FLIR sequences. Motion is a very strong cue, especially in highly cluttered environments, that has not been considered sufficiently in previous work. The shortcomings of the sensor and requirements for real-time processing induce the need for a fast and robust system. Our detection system adapts well-known robust techniques from the visible to the FLIR domain. An iterative approach, used for the most time-costly operation, image registration, assures a scalable algorithm complexity. We propose a new methodology to link the new dynamic information and static cues, such as object pose, enabling the construction of more redundant and fault-tolerant systems. Our algorithm has been implemented, and results on difficult, real sequences are presented.

In future work, we want to integrate the presented dynamic scene analysis system with existing static image ATR systems (such as [NA96]) into a comprehensive system (Fig. 12). The shaded box highlights the parts of the system described in this paper. Together with cues from other modules, it can be used in a Bayesian sensor fusion paradigm to improve detection accuracy and reduce false alarms. In such a fusion stage detection, recognition and pose results

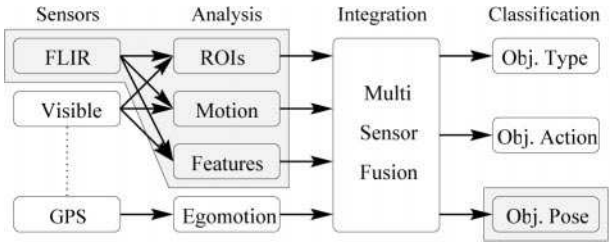


Fig. 12. Integration of MODEEP (*shaded*) into a future target detection and recognition system

from various cues such as motion, target shape, size or parts can be integrated using a Bayesian meta-classifier. The different paradigms can be used to mutually verify their results and synergetically improve performance. Compared to existing systems, dynamic scene analysis enables the inclusion of target action recognition. This action recognition enables multi-frame descriptions such as object 'starts' and 'stops' and 'changes in acceleration' and 'changes in direction' to be extracted automatically. Target action knowledge provides a high-level abstraction based on motion analysis that has great potential to extend and enhance existing systems.

References

[Adi85] Adiv G (1985) Determining three-dimensional motion and structure from optical flow generated by several moving objects. IEEE Trans Pattern Anal Mach Intell 7(4): 384–400

[AN89] Aggarwal JK, Nandhakumar N (1989) On the computation of motion from sequences of images: A review. Proc IEEE 76: 917–935

[BA83] Burt P, Adelson EH (1983) The Laplacian pyramid as a compact image code. IEEE Trans Commun 31(4): 532–540

[BA96] Black MJ, Anandan P (1996) The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. Comput Vision Image Understanding 63(1): 75–104

[BAHH92] Bergen JR, Anandan P, Hanna KJ, Hingorani R (1992) Hierarchical model-based motion estimation. In: Sandini G (ed) Proceedings European Conference on Computer Vision (LNCS 588), 1992, Berlin, Germany. Springer, Berlin Heidelberg New York, pp 237–252

[BAK91] Battiti R, Amaldi E, Koch C (1991) Computing optical flow across multiple scales: an adaptive coarse- to-fine strategy. Int J Comput Vision 6(2): 133–145

[BB90] Burger W, Bhanu B (1990) Estimating 3-D egomotion from perspective image sequences. IEEE Trans Pattern Anal Mach Intell 12(11): 1040–1058

[BDZ+97] Bhanu B, Dudgeon DE, Zelnio EG, Rosenfeld A, Casasent D, Reed IS (1997) Introduction to the special issue on automatic target detection and recognition. IEEE Trans Image Process 6(1): 1–6

[Can86] Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8(6): 679– 698

[DH73] Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York, N.Y.

[FvDFH90] Foley JD, Dam A van, Feiner SK, Hughes JF (1990) Computer graphics: Principles and practice. Addison- Wesley, Reading, Mass.

[HS81] Horn BKP, Schunk BG (1981) Determining optical flow. Artif Intell 17: 185–203

[Hub81] Huber PJ (1981) Robust statistics. Wiley, New York, N.Y.

[IA98] Irani M, Anandan P (1998) A unified approach to moving-object detection in 2D and 3D scenes. IEEE Trans Pattern Anal Mach Intell 20(6): 577–589

[MDD+95] Morimoto CH, Dementhon D, Davis LS, Chellappa R, Nelson R (1995) Detection of independently moving objects in passive video. In: Masaky I (ed) Proceedings of Intelligent Vehicles Workshop, September 1995, Detroit, Mich. IEEE, pp 270–275

[NA92] Nandhakumar N, Aggarwal JK (1992) Multisensory computer vision. Adv Comput 60: 34

[NA96] Nair D, Aggarwal JK (1996) A focused target segmentation paradigm. In: Cipolla R, Buxton B (eds) Fourth European Conference on Computer Vision, April 1996, Cambridge, UK. Springer, Berlin Heidelberg New York, pp 579–588

[RCM+95] Rogers SK, Colombi JM, Martin CE, Gainey JC, Fielding KH, Burns TJ, Ruck DW, Kabrisky M, Oxley M (1995) Neural networks for automatic target recognition. Neural Networks 8(7/8): 1153–1184

[RWBG97] Ratches JA, Walters CP, Buser RG, Guenther BD (1997) Aided and automatic target recognition based upon sensory inputs from image forming systems. IEEE Trans Pattern Anal Mach Intell 19(9): 1004–1019

[Sob70] Sobel IE (1970) Camera models and machine perception. Ph.D. thesis. Stanford University, Stanford, CA

[TZ98] Torr PHS, Zisserman A (1998) Concerning Bayesian motion segmentation, model averaging, matching and the trifocal tensor. In: Burkhardt H, Neumann B (eds) Fifth European Conference on Computer Vision, June 1998, Freiburg, Germany. Springer, Berlin Heidelberg New York, pp 511–527

[WA94] Wang JYA, Adelson EH (1994) Spatio- temporal segmentation of video data. Proc SPIE (Image and Video Processing II) 2182: 120–131

[Wan95] Wandell BA (1995) Foundations of vision. Sinauer Associates, Sunderland, MA

[Wu95] Wu QX (1995) A correlation-relaxation- labeling framework for computing optical flow- template matching from a new perspective. IEEE Trans Pattern Anal Mach Intell 17(9): 843–853

Alexander Strehl received his Vordiplom in computer science from the Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, in 1996 and the M.Sc.Eng. degree in electrical and computer engineering from The University of Texas at Austin in 1998. During his studies he worked as a research assistant for the Fraunhofer Gesellschaft and as a management consultant for McKinsey & Company, Inc. He is presently working towards the Ph.D. degree at The University of Texas at Austin. Alexander Strehl is a member of IEEE and Phi Kappa Phi. His research interests include computer vision, video processing, large scale data mining, collaborative filtering, clustering, permission marketing, and e-commerce.

J. K. Aggarwal has served on the faculty of The University of Texas at Austin College of Engineering since 1964 and is currently the Cullen Professor of Electrical and Computer Engineering and Director of the Computer and Vision Research Center. His research interests include computer vision and and pattern recognition. A Fellow of IEEE since 1976 and IAPR since 1998, he received the Senior Research Award of the American Society of Engineering Education in 1992, and the 1996 Technical Achievement Award of the IEEE Computer Society. He is author or editor of 7 books and 38 book chapters; author of over 175 journal papers, as well as numerous proceedings papers and technical reports. He has served as Chairman of the IEEE Computer Society Technical Committee on Pattern Analysis and Machine Intelligence (1987–1989); Director of the NATO Advanced Research Workshop on Multisensor Fusion for Computer Vision, Grenoble, France (1989); Chairman of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1993), and President of the International Association for Pattern Recognition (1992-94). He currently serves as IEEE Computer Society representative to the IAPR.