Sensors **2014**, *14*, 10124-10145; doi:10.3390/s140610124

OPEN ACCESS
sensors
ISSN 1424-8220
www.mdpi.com/journal/sensors

Article

Joint Target Tracking, Recognition and Segmentation for Infrared Imagery Using a Shape Manifold-Based Level Set

Jiulu Gong [1], Guoliang Fan [2,*], Liangjiang Yu [2], Joseph P. Havlicek [3], Derong Chen [1] and Ningjun Fan [1]

[1] School of Mechatronical Engineering, Beijing Institute of Technology, No. 5, Zhongguancun South Street, Haidian District, Beijing 100081, China; E-Mails: gongjiulu@gmail.com (J.G.); cdr@bit.edu.cn (D.C.); njfan@bit.edu.cn (N.F.)

[2] School of Electrical and Computer Engineering, Oklahoma State University, 202 Engineering South, Stillwater, OK 74078, USA; E-Mail: liangjiang.yu@okstate.edu

[3] School of Electrical and Computer Engineering, University of Oklahoma, 110 West Boyd, DEH 150 Norman, OK 73019, USA; E-Mail: joebob@ou.edu

* Author to whom correspondence should be addressed; E-Mail: guoliang.fan@okstate.edu; Tel.: +1-405-744-1547; Fax: +1-405-744-9198.

Received: 24 March 2014; in revised form: 22 May 2014 / Accepted: 23 May 2014 / Published: 10 June 2014

Abstract: We propose a new integrated target tracking, recognition and segmentation algorithm, called ATR-Seg, for infrared imagery. ATR-Seg is formulated in a probabilistic shape-aware level set framework that incorporates a joint view-identity manifold (JVIM) for target shape modeling. As a shape generative model, JVIM features a unified manifold structure in the latent space that is embedded with one view-independent identity manifold and infinite identity-dependent view manifolds. In the ATR-Seg algorithm, the ATR problem formulated as a sequential level-set optimization process over the latent space of JVIM, so that tracking and recognition can be jointly optimized via implicit shape matching where target segmentation is achieved as a by-product without any pre-processing or feature extraction. Experimental results on the recently released SENSIAC ATR database demonstrate the advantages and effectiveness of ATR-Seg over two recent ATR algorithms that involve explicit shape matching.

Keywords: automatic target recognition; joint tracking recognition and segmentation; shape manifolds; level set; manifold learning

1. Introduction

As a challenging problem in pattern recognition and machine learning for decades, automatic target tracking and recognition (ATR) has been an important topic for many military and civilian applications. Infrared (IR) ATR is a more challenging problem due to two main reasons. First, an IR target's appearance may change dramatically under different working conditions and ambient environment. Second, the IR imagery usually has poor quality compared with the visible one. There are two important and related research issues in ATR research, appearance representation and motion modeling [1]. The former one focuses on capturing distinct and salient features (e.g., edge, shape, texture) of a target, and the latter one tries to predict the target's state (e.g., position, pose, velocity) during sequential estimation. They could play a complementary role in an ATR process [2].

Shape is a simple yet robust, feature for target representation in many ATR applications. There are three commonly used ways of shape representation: a 3D mesh model [3], 2D shape templates [4,5] and a manifold-based shape generative model learned from 2D snapshots [6–8]. When a 3D model was used, a 3D-to-2D projection is needed to get the 2D shapes according to the camera model and the target's position. Using a 3D model for shape modeling usually needs more memory and expensive computational resources. In [5], a 2D shape template was used to represent the target's appearance, and an online learning was used to update this shape model under different views. Manifold learning methods have proven to be powerful for shape modeling by providing a variety of meaningful shape prior to assist or constrain the shape matching process. In [8], a couplet of view and identity manifolds (CVIM) was proposed for multi-view and multi-target shape modeling, where target pre-segmentation was implemented via background subtraction and the ATR inference involves explicit shape matching between segmented targets and shapes hypothesis generated by CVIM.

In this work, we propose a new particle filter-based ATR-Seg (segmentation) algorithm that integrates JVIM (joint view-identity manifold) with a shape-aware level set energy function which leads to a joint tracking, recognition and segmentation framework. JVIM encapsulates two shape variables, identity and view, in a unified latent space, which is embedded with one view-independent identity manifold and infinite identity-dependent view manifolds. Unlike CVIM obtained via nonlinear tensor decomposition, JVIM is learned via a modified Gaussian process latent variable model [9] which leads to a probabilistic shape model. Also, a stochastic gradient descent method [10] is developed to speed up JVIM learning, and a local approximate method is used for fast shape interpolation and efficient shape inference. Furthermore, we integrate JVIM with a level set energy function that is able to evaluate how likely a shape synthesized by JVIM can segment out a valid target from an image. This energy function is adopted as the likelihood function in the particle filter where a general motion model is used for handling highly maneuverable targets. The performance of ATR-Seg was evaluated using the SENSIAC (Military Sensing Information Analysis Center) IR dataset [11], which demonstrated the advantage of the proposed method over several methods that involve target pre-segmentation and explicit shape matching.

The remainder of this paper is organized as follow. In Section 2, we review some related works on shape manifold learning and shape matching. In Section 3, we use a graphical model to develop a probabilistic framework of our ATR-Seg algorithm. In Section 4, we introduce JVIM for general shape modeling. In Section 5, we present a shape-aware level set energy function for implicit shape matching.

In Section 6, we present a particle filter-based sequential inference method for ATR-Seg. In Section 7, we evaluate the proposed ATR-Seg algorithm in two aspects, *i.e.*, JVIM-based shape modeling and implicit shape matching which are involved in the likelihood function of the particle filter. We conclude our paper in Section 8.

2. Related Works

ATR itself is a broad field involving diverse topics. Due to the fact that shape modeling is the key issue in our ATR research, our review below will be focused on two shape-related topics, manifold-based shape modeling and shape matching.

2.1. Manifold-Based Shape Modeling

A manifold-based shape model can be learned from a set of exemplar shapes and is able to interpolate new shapes from the low-dimensional latent space. Roughly speaking, there are three manifold learning approaches for shape modeling, geometrically-inspired methods, latent variable models, and hybrid models. The first approach seeks to preserve the geometric relationships among the high-dimensional data in the low-dimensional space, e.g., IsoMap [12], Local Linear Embedding (LLE) [13], Diffusion Maps [14] and Laplacian Eigenmaps [15]. These methods focus on how to explore the geometric structure among the high-dimensional data and how to maintain this structure in the low dimensional embedding space. However, the mapping relationship from the latent space and the data space is not available and has to be learned separately. The second approach represents the shape data by a few latent variables along with a mapping from the latent space to the data space, such as PCA [16], PPCA [17], KPCA [18], Gaussian Process Latent Variable Models (GPLVM) [19] and tensor decomposition [20], *et al*. GPLVM is a probabilistic manifold learning method which employs the Gaussian process as the nonlinear mapping function. Above approaches are data driven shape modeling methods without involving prior knowledge in the latent space, and as a result, the shape-based inference process may be less intuitive due to the lack of a physically meaningful manifold structure.

To support a more meaningful and manageable manifold structure while preserving the mapping function, there is a trend to combine the first two approaches along with some topology prior for manifold learning [21]. In [9], the local linear GPLVM (LL-GPLVM) was proposed for complex motion modeling, which incorporates a LLE-based topology prior in the latent space. Specifically, a circular-shaped manifold prior is used to jointly model both "walking" and "running" motion data in a unified cylinder-shaped manifold. In [8], CVIM was proposed for shape modeling via nonlinear tensor decomposition where two independent manifolds, an identity manifold and a view manifold, were involved. Specifically, the view manifold was assumed to be a hemisphere that represents all possible viewing angles for a ground target, and the identity manifold was learned from the tensor coefficient space that was used to interpolate "intermediate" or "unknown" target types from known ones. A key issue about the identity manifold is the determination of manifold topology, *i.e.*, the ordering relationship across all different target types. Sharing a similar spirit of IsoMap, the shortest-closed-path is used to find the optimal manifold topology that allows targets with similar shapes to stay closer and those with

dissimilar shapes far away. This arrangement ensures the best local smoothness and global continuity that are important for valid shape interpolation along the identity manifold.

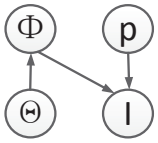
2.2. Shape Matching

In shape-based tracking algorithms, there are two ways to measure shape similarity: explicit shape matching and implicit shape matching. The former one involves a direct spatial comparison between two shapes, an observed one and a hypothesized one, by using a certain distance metric. In such a case, pre-processing or feature extraction, e.g., background subtraction in [8], is needed prior to tracking and recognition, which is relatively manageable for a stationary sensor platform and may need additional computational load in a general case. Moreover, the overall ATR performance could be sensitive to the pre-processing results. The latter one represents a shape implicitly by a level set embedding function which can be used to evaluate the segmentation quality of a given shape in an image. For example, a shape-constrained energy function was used in [6,7] to evaluate how likely the given shape can segment out a valid object, where a gradient descent method was used to optimize this energy function to achieve tracking and segmentation jointly. Therefore, implicit shape matching does not involve any pre-processing or feature extraction beforehand, however, due to the lack of dynamic modeling in level set optimization, it is still hard to track highly maneuverable targets by the traditional data-driven gradient descent optimization method. As pointed in [22], motion/dynamic modeling is an important step for most ATR applications. This motivates our research to augment a motion model in implicit shape matching for maneuverable target tracking.

3. ATR-Seg Problem Formulation

We list all symbols used in this paper in Table 1. Given the observed video sequence $\mathbf{I}_t$, with $t = 1, \ldots, T$, where T is the total number of image frames, the objective of ATR-Seg is to (1) find the 3D position of a target in the camera coordinate $\mathbf{p}$ (tracking) or 2D image coordinate, (2) to identify the target type α (recognition), along with the view angle φ (pose estimation), and (3) to segment the target-of-interest that best explains the observation data $\mathbf{\Phi}$ (segmentation). The 2D shape of a target can be determined by the target type α, and view angle φ, so we define $\mathbf{\Theta} = [\alpha, \varphi]$ to represent two shape related variables. The conditional dependency among all variables is shown in Figure 1.

Figure 1. Graphical modeling for the proposed ATR-Seg algorithm, where $\mathbf{I}_t$ represents an image frame, $\mathbf{p}$ 3D target position, $\mathbf{\Phi}$ target segmentation, and $\mathbf{\Theta}$ the set of shape variables.



According to Figure 1, we define the objective function of ATR-Seg from the joint distribution $p(\mathbf{p}_t, \mathbf{\Theta}_t, \mathbf{\Phi}, \mathbf{I}_t)$ which can be written (t is omitted for simplicity) as:

$$p(\mathbf{p}, \mathbf{\Theta}, \mathbf{\Phi}, \mathbf{I}) = p(\mathbf{I}|\mathbf{p}, \mathbf{\Phi})p(\mathbf{\Phi}|\mathbf{\Theta})p(\mathbf{\Theta})p(\mathbf{p}) \quad (1)$$

Table 1. Descriptions of all mathematical symbols.

Symbols used in Problem Formulation (Section 3)	
$\mathbf{p}$	the target's 3D position $\mathbf{p} = [p_x, p_y, p_z]^T$
α	the target type
φ	the view angle
$\mathbf{\Theta}$	shape related variables ($[\alpha, \varphi]$)
$\mathbf{\Phi}$	a target shape segmentation
$\mathbf{I}_t$	an observed image frame at time t
Symbols used in JVIM-based shape modeling (Section 4)	
$\mathbf{Y}$	JVIM training data
$\mathbf{X}$	JVIM latent space
θ	the aspect angle of a target
ϕ	the elevation angle of a target
β	the kernel hyper-parameters of JVIM
d	the dimension of the shape space
$\mathbf{w}$	the LLE coefficients for local topology encoding
L_{JVIM}	the JVIM objective function
L_D	the data term in L_{JVIM}
L_T	the topology term in L_{JVIM}
$\mathbf{K}_Y$	the covariance matrix of JVIM learning
$\mathbf{x}_r$	a reference latent point in JVIM learning
$\mathbf{X}_R$	the neighborhood of $\mathbf{x}_r$ for local learning
M_1	the size of $\mathbf{X}_R$ (the range of local learning)
$\mathbf{Y}_R$	the corresponding shape for $\mathbf{X}_R$
N	the size of training data
$\mathbf{x}'$	a new latent point for JVIM-based shape interpolation
$\mathbf{X}'$	the neighborhood of $\mathbf{x}'$ for local inferencing
M_2	the size of $\mathbf{X}'$ (the range of local inferencing)
$\mathbf{Y}'$	the corresponding shape data for $\mathbf{X}'$
$k(\mathbf{x}_1, \mathbf{x}_2)$	a RBF kernel function in JVIM
$\hat{\boldsymbol{\mu}}_{x'}$	an interpolated shape at $\mathbf{x}'$ via JVIM
$\hat{\sigma}^2_{x'}$	uncertainty of shape interpolation at $\mathbf{x}'$
Symbols used in shape-aware level set (Section 5)	
x	a 2D pixel location in an image frame
y	a pixel intensity value
M	foreground/background models $M = \{M_f, M_b\}$
$H_\epsilon[\cdot]$	the smoothed Heaviside step function
Symbols used in sequential inference (Section 6)	
ψ_t	the heading direction of a ground vehicle in frame t
v_t	the target velocity along ψ_t in frame t
Δt	the time interval of two adjacent frames
$\mathbf{Z}_t$	the state vector in frame t ($\mathbf{Z}_t = [\mathbf{p}_t^T, v_t, \psi_t, \alpha_t]^T$)

By using the Bayesian theorem, we can get the posterior as:

$$p(\mathbf{p}, \boldsymbol{\Theta}, \boldsymbol{\Phi}|\mathbf{I}) \propto p(\mathbf{I}|\mathbf{p}, \boldsymbol{\Phi})p(\boldsymbol{\Phi}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})p(\mathbf{p}) \tag{2}$$

which encapsulates three major components in the proposed ATR-Seg algorithm, as shown below:

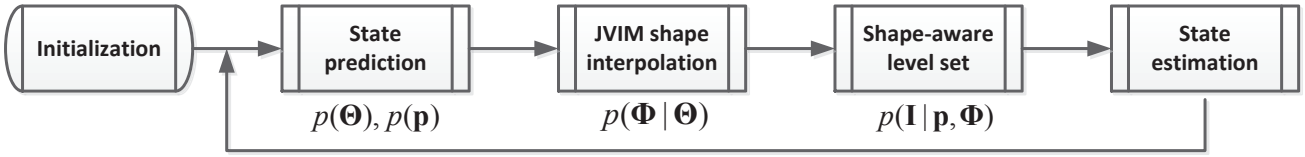
- Shape manifold learning provides a mapping from $\boldsymbol{\Theta}$ to $\boldsymbol{\Phi}$, *i.e.*, $p(\boldsymbol{\Phi}|\boldsymbol{\Theta})$. In Section 4, JVIM is proposed for multi-view and multi-target shape modeling, which features a novel manifold structure with one view-independent identity manifold and infinite identity-dependent view manifolds to impose a conditional dependency between the two shape-related factors, view and identity, in a unified latent space.

- Shape-aware level set $p(\mathbf{I}|\mathbf{p}, \boldsymbol{\Phi})$ measures how likely $\boldsymbol{\Phi}$ can segment a valid target at position $\mathbf{p}$ in image I. In Section 5, a shape-aware level set energy function is proposed for implicit shape matching, which evaluates the segmentation quality.

- Pose/position priors $\boldsymbol{\Theta}$ and $\mathbf{p}$, *i.e.*, $p(\boldsymbol{\Theta})$ and $p(\mathbf{p})$, *i.e.*, are important to track highly manoeuverable targets in a sequential manner. In Section 6, sequential shape inference method is presented that involve dynamic priors for $\boldsymbol{\Theta}$ and $\mathbf{p}$ using a 3D motion model.

The flowchart for ATR-Seg is shown in Figure 2 where four steps are involved sequentially and recursively. First, state prediction will draw a set of samples to predict all state variables (position/angle/identity). Second, a series of shape hypotheses are created via JVIM in some hypothesized locations according to predicted state information. Third, a level-set energy function is used as the likelihood function to weight each hypothesized shape/location that quantifies how well that shape can segment a valid target in that particular location. Fourth, state estimation at the current frame is obtained by the conditional mean of all weighted samples and will be used for state prediction in the next frame.

Figure 2. Flowchart for ATR-Seg.

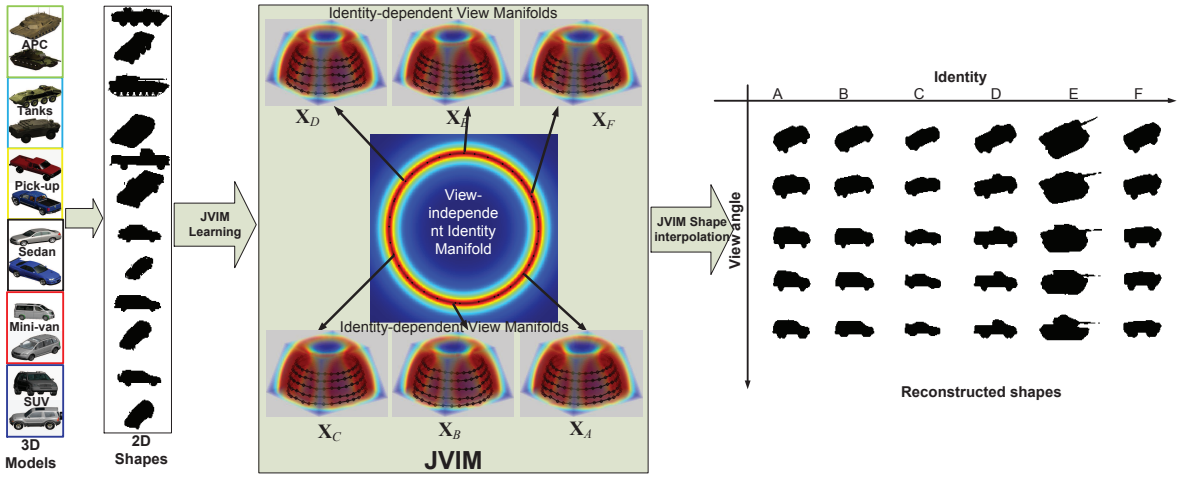


4. Joint View-Identity Manifold (JVIM)

JVIM is learned from a set of 2D shape exemplars $\mathbf{Y}$ generated from a set of 3D CAD models. The latent space $\mathbf{X}$ can be represented by two variables, identity α and view φ (including aspect angle θ and elevation angle ϕ), which are defined along their respective manifolds. Considering the fact that all targets have different 3D structures, leading to different view manifolds, and they keep the same identity under different views, we impose a conditional dependency between α and φ in JVIM that encapsulates one view-independent identity manifold and infinite identity-dependent view manifolds. Specifically,

the identity manifold represents the view-independent shape variability across different target types, and an identity-specific view manifold captures the shape variability of a target under different views. Motivated by [8,23], the identity manifold is simplified to have a circular-shaped topology prior, which facilitates manifold learning and shape inference. Intuitively, a hemispherical-shaped topology prior is assumed for identity-specific view manifold, which represents all possible aspect and elevation angles for ground vehicle. All topology priors are encoded by LLE and incorporated into the GPLVM-based learning framework, as shown in Figure 3.

Figure 3. JVIM learning and shape interpolation, where one view-independent identity manifold and six identity-dependent view manifolds are color-coded according to the uncertainty of GP mapping. (Adapted from [24], with permission from Elsevier.)



The objective of JVIM learning is to find $\mathbf{X}$ and β by maximizing $p(\mathbf{Y}|\mathbf{X}, \beta, \mathbf{w})$, where β is the mapping parameter and $\mathbf{w}$ represents the LLE-based topology prior in the latent space. The Gaussian process (GP) is used as the nonlinear mapping function from the latent space to the shape space ($\mathbf{X} \rightarrow \mathbf{Y}$), and the objective function of JVIM learning is written as:

$$p(\mathbf{Y}|\mathbf{X}, \beta, \mathbf{w}) = p(\mathbf{Y}|\mathbf{X}, \beta)p(\mathbf{X}|\mathbf{w}) \tag{3}$$

where:

$$p(\mathbf{Y}|\mathbf{X}, \beta)\frac{1}{\sqrt{(2\pi)^{Nd}|\mathbf{K}_Y|^d}}\exp(-\frac{1}{2}tr(\mathbf{K}_Y^{-1}\mathbf{Y}\mathbf{Y}^T)) \tag{4}$$

where d is the dimension of the shape space and β denotes the kernel hyper-parameters used in the covariance matrix, $\mathbf{K}_Y$. It is worth noting that Equation (4) is similar to the objective function of GPLVM [19], and:

$$p(\mathbf{X}|\mathbf{w}) = \frac{1}{Z}\exp\left(-\frac{1}{\sigma^2}\sum_{i=1}^{N}\|\mathbf{X}_i - \sum_{j=1}^{N}\mathbf{w}_{ij}\mathbf{X}_j\|^2\right) \tag{5}$$

where $\mathbf{w}$ is the set of LLE weights to reconstruct each latent point from its local neighboring points by minimizing $f(\mathbf{w}) = \sum_{i=1}^{N}\|\mathbf{X}_i - \sum_{j=1}^{N}\mathbf{w}_{ij}\mathbf{X}_j\|^2$, Z is a normalization constant, σ^2 represents a global scaling of the prior and N the number of training samples. Furthermore, the negative log operation is used to simplify the objective function as:

$$L_{JVIM} = -\log p(\mathbf{Y}|\mathbf{X}, \beta)p(\mathbf{X}|\mathbf{w}) = L_D + L_T + C \tag{6}$$

where C is a constant.

JVIM learning involves a gradient descent method to minimize the objective function defined in Equation (3) with respect to $\mathbf{X}$ and β. With an $O(N^3)$ operation required at each iteration, it is computationally prohibitive for a large training data set. The stochastic gradient descent proposed in [10] is adapted to be a local updating according to the unique structure of JVIM to approximate the gradients locally. At each iteration, the reference point, $\mathbf{x}_r$, is chosen randomly, and the derivatives w.r.t $\mathbf{X}_R$ and β are calculated as:

$$\frac{\partial L_D}{\partial \mathbf{X}_R} \approx -(\mathbf{K}_R^{-1}\mathbf{Y}_R\mathbf{Y}_R^T\mathbf{K}_R^{-1} - d\mathbf{K}_R^{-1}) \cdot \frac{\partial \mathbf{K}_R}{\partial \mathbf{X}_R} \tag{7}$$

$$\frac{\partial L_{JVIM}}{\partial \beta} \approx -(\mathbf{K}_R^{-1}\mathbf{Y}_R\mathbf{Y}_R^T\mathbf{K}_R^{-1} - d\mathbf{K}_R^{-1}) \cdot \frac{\partial \mathbf{K}_R}{\partial \beta} \tag{8}$$

where $\mathbf{X}_R$ is the neighborhood for a reference point, $\mathbf{x}_r$, of size M_1, $\mathbf{Y}_R$ is the corresponding shape data and $\mathbf{K}_R$ ($M_1 \times M_1$) is the kernel matrix of $\mathbf{X}_R$. The neighborhood for each training data can be pre-assigned according to the topology structure, and the gradients are estimated stochastically, locally and efficiently.

As a generative model, given an arbitrary latent point in $\mathbf{X}$, JVIM can generate the corresponding shape via Gaussian Process (GP) mapping. For real-time applications, shape interpolation must be carried out efficiently, which is difficult for a large training data set with high dimensionality. Inspired by [25], a GP can be approximated by a set of local GPs, in JVIM-based shape interpolation, the kernel matrix is computed locally from a set of training data that are close to the given point. Given $\mathbf{x}'$, we first find its closest training point, which has a pre-assigned neighborhood, $\mathbf{X}'$, of size M_2; then, $\mathbf{X}'$ and the corresponding shape data $\mathbf{Y}'$ are used to approximate the mean and variance of GP mapping as:

$$\hat{\boldsymbol{\mu}}_{x'} = \mathbf{k}_{x'X'}^T\mathbf{K}_{Y'}^{-1}\mathbf{Y}' \tag{9}$$

$$\hat{\sigma}_{x'}^2 = k(\mathbf{x}', \mathbf{x}') - \mathbf{k}_{x'X'}^T\mathbf{K}_{Y'}^{-1}\mathbf{k}_{x'X'} \tag{10}$$

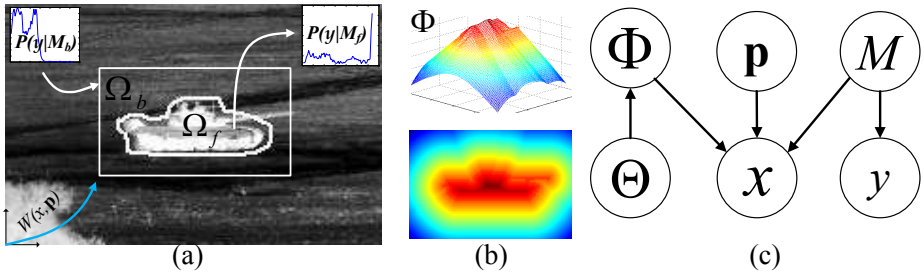
where $\mathbf{k}_{x'X'}$ is a vector made of $k(\mathbf{x}', \mathbf{x}_i)$ ($\mathbf{x}_i \in \mathbf{X}'$) and $\mathbf{K}_{Y'}$ ($M_2 \times M_2$) is the local covariance matrix computed from $\mathbf{X}'$. More detail about JVIM learning and inference can be found in our previous work [26], where explicit shape matching is involved. In the following section, we will introduce implicit shape matching by incorporating a shape-aware level set for target tracking and recognition, where target segmentation becomes a by-product.

5. Shape-Aware Level Set

JVIM is used to provide a useful shape prior that can be further combined with the level set to define an energy function for implicit shape matching. This is called the shape-aware level set, which does not involve feature extraction or target pre-segmentation. The shape-aware level set in this work is distinct from that in [6,7,27] primarily in two aspects. Firstly, the shape generated model in [6,7], which was less structured with little semantic meaning and, was limited to object recognition/segmentation under

the same view or human pose estimation for the same person along the same walking path. JVIM is a multi-view and multi-target shape model that has a well-defined semantic structure, which supports robust ATR for different targets under arbitrary view angles. Secondly, a gradient decent method was used for level set optimization in [6,7,27], which does not involve a motion model and makes it hard to track highly maneuverable targets. In this work, a 3D motion model is used to combine the position/pose priors into a sequential optimization model to improve the robustness and accuracy of ATR-Seg.

Figure 4. Shape-aware level set model for implicit shape matching. (**a**) Illustration of a target in an infrared image: foreground Ω_f and background Ω_b, foreground/background intensity models M, and the 3D-2D camera projection W(x,$\mathbf{p}$). (**b**) The shape embedding function $\mathbf{\Phi}$. (**c**) The graphical model for shape-aware level set, where $\mathbf{p}$ is the target 3D location of a ground-vehicle, and $\mathbf{\Theta}$ is the shape parameter in JVIM.



(a) (b) (c)

As shown in Figure 4a, we represent an image by $\mathbf{I} = \{x_i, y_i\}$, where $1 \leq i \leq n$, n is the number of pixels in $\mathbf{I}$ and x and y are the pixel 2D location and pixel intensity value, respectively. We introduce a parameter, M, to represent the foreground/background models $M = \{M_f, M_b\}$; then, the original graphical model of ATR-Seg in Figure 1 will become the one in Figure 4c. which defines a joint distribution of all parameters for each pixel (x_i, y_i) as

$$p(x_i, y_i, \mathbf{p}, \mathbf{\Theta}, \mathbf{\Phi}, M) = p(x_i|\mathbf{p}, \mathbf{\Phi}, M)p(y_i|M)p(\mathbf{\Phi}|\mathbf{\Theta})p(M)p(\mathbf{\Theta})p(\mathbf{p}) \tag{11}$$

where $\mathbf{\Phi}$ is a shape represented by the level set embedding function shown in Figure 4b and $p(\mathbf{\Phi}|\mathbf{\Theta})$ corresponds to JVIM-based shape interpolation via GP mapping. A histogram is used for foreground/background appearance model $p(y_i|M)$, and the number of bins is dependent on the size of the target and gray scale. In order to get the posterior, $p(\mathbf{p}, \mathbf{\Theta}, \mathbf{\Phi}, M|x_i, y_i)$, which will be used to develop the objective function for ATR-Seg, we take the same strategy as in [27]. First, divide Equation (11) by $p(y_i) = \sum_{j\in\{f,b\}} p(y_i|M_j)p(M_j)$:

$$p(x_i, \mathbf{p}, \mathbf{\Theta}, \Phi, M|y_i) = p(x_i|\mathbf{p}, \mathbf{\Phi}, M)p(M|y_i)p(\mathbf{\Phi}|\mathbf{\Theta})p(\mathbf{\Theta})p(\mathbf{p}) \tag{12}$$

where $p(M|y_i)$ is given by:

$$p(M_j|y_i) = \frac{p(y_i|M_j)p(M_j)}{\sum_{k\in\{f,b\}} p(y_i|M_k)p(M_k)}, \quad j \in \{f, b\} \tag{13}$$

Upon dividing Equation (12) by $p(x_i) = 1/n$ and marginalizing over the models, M, we obtain:

$$p(\mathbf{p}, \mathbf{\Theta}, \mathbf{\Phi}|x_i, y_i) = n \sum_{j\in\{f,b\}} p(x_i|\mathbf{p}, \mathbf{\Phi}, M_j)p(M_j|y_i)p(\mathbf{\Phi}|\mathbf{\Theta})p(\mathbf{\Theta})p(\mathbf{p}) \tag{14}$$

Assuming all pixels are independent, the posterior for all pixels in a frame is then given by:

$$\begin{aligned} p(x_i|\mathbf{p},\mathbf{\Phi},M_f) &= H_\epsilon[\mathbf{\Phi}(\mathbf{x}_i)]/\eta_f \\ p(x_i|\mathbf{p},\mathbf{\Phi},M_b) &= \{1-H_\epsilon[\mathbf{\Phi}(x_i)]\}/\eta_b \end{aligned} \tag{15}$$

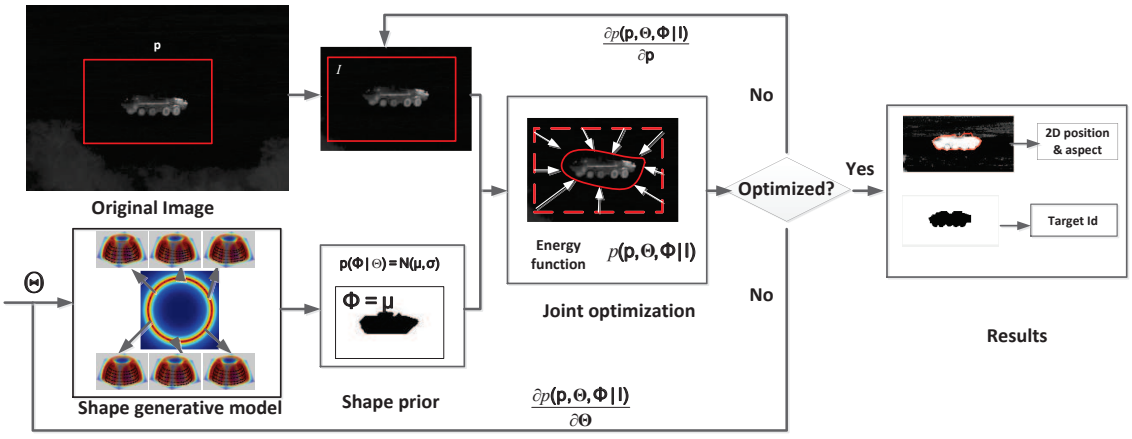
where $H_\epsilon[\cdot]$ is the smoothed Heaviside step function, $\eta_f = \sum_{i=1}^{n} H_\epsilon[\mathbf{\Phi}(x_i)]$, $\eta_b = \sum_{i=1}^{n}\{1-H_\epsilon[\mathbf{\Phi}(x_i)]\}$ and $p(M_j) = \eta_j/n$ for $j \in \{f,b\}$.

Then, from Equations (2), (14) and (15), we have:

$$p(\mathbf{I}|\mathbf{p},\mathbf{\Phi}) \propto \prod_{i=1}^{n} \sum_{j\in\{f,b\}} p(x_i|\mathbf{p},\mathbf{\Phi},M_j)p(M_j|y_i) \tag{16}$$

which evaluates how likely shape $\mathbf{\Phi}$ can segment a valid target from $\mathbf{I}$ at position $\mathbf{p}$. The objective function in Equation (2) can be optimized through a gradient descent method similar to the one in [7], which is illustrated in Figure 5. As shown in Figure 5, JVIM is firstly used to generate a shape hypothesis, $\mathbf{\Phi}^0$, given initial identity and view angle $\mathbf{\Theta}^0$; then, $\mathbf{\Phi}^0$ is used to initialize the objective function, $p(\mathbf{p},\mathbf{\Theta},\mathbf{I})$, for initial position, $\mathbf{p}^0$. We take the derivative of $p(\mathbf{p},\mathbf{\Theta},\mathbf{\Phi}|\mathbf{I})$ with respect to $\mathbf{\Theta}$ and $\mathbf{p}$ to get $\frac{\partial p(\mathbf{p},\mathbf{\Theta},\mathbf{\Phi}|\mathbf{I})}{\partial \mathbf{\Theta}}$ and $\frac{\partial p(\mathbf{p},\mathbf{\Theta},\mathbf{\Phi}|\mathbf{I})}{\partial \mathbf{p}}$, which will be used to update $\mathbf{\Theta}$ and $\mathbf{p}$ until the objective function converges. When $p(\mathbf{p},\mathbf{\Theta},\mathbf{\Phi}|\mathbf{I})$ is maximized, we output the updated target's 2D position, $\mathbf{p}^*$, target identity and view angle $\mathbf{\Theta}^*$, as well as the updated shape $\mathbf{\Phi}^*$ that can best segment the image.

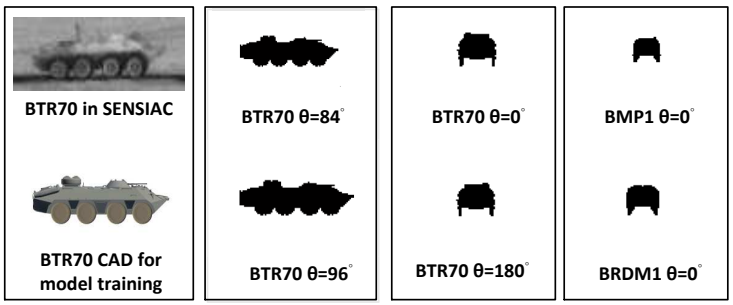
Figure 5. Optimization of ATR-Seg by a gradient descent method.



This method works well on a single image when a good initialization is given in the latent space of JVIM. However, it may fail quickly when dealing with an image sequence with a highly maneuverable target, due to four possible cases of shape ambiguity, as shown in Figure 6, which makes data-driven optimization not reliable in practice. (1) The first is due to the possible shape mismatch between the CAD models and real targets, even for the same target type (Figure 6a). (2) The second is due to the symmetry property of a target's shape (Figure 6b), which means a target may present a similar shape at different (e.g., supplement) aspect angles, especially when the elevation angle is zero (Figure 6b). (3) The third is due to the ambiguity of the front/rear views when a target looks very similar (Figure 6c). (4) The fourth is similar to the previous one in which many targets look alike at the front/rear views

(Figure 6d). These factors make the gradient-based approach not effective at dealing with a maneuvering target. A possible remedy is to introduce a dynamic motion model to support robust sequential shape inference based on JVIM, as to be discussed below.

Figure 6. Possible reasons for the failure of the gradient descent method.



(a) Shape ambiguity-I (model-data mismatching) (b) Shape ambiguity-II (symmetry views) (c) Shape ambiguity-III (front/rear views) (d) Shape ambiguity-IV (shape similarity)

6. Sequential Shape Inference

Essentially, the objective of ATR-Seg is to perform sequential shape inference from an image sequence by maximizing the posterior of $p(\mathbf{p}_t, \mathbf{\Theta}_t, \mathbf{\Phi}_t|\mathbf{I}_t)$. According to Figure 1 in Section 3, $\mathbf{\Phi}$ is only dependent on $\mathbf{\Theta}$, so the objective function can be rewritten as:

$$p(\mathbf{p}_t, \mathbf{\Theta}_t, \mathbf{\Phi}_t|\mathbf{I}_t) = p(\mathbf{p}_t, \mathbf{\Theta}_t|\mathbf{I}_t)p(\mathbf{\Phi}_t|\mathbf{\Theta}_t) \tag{17}$$

where $p(\mathbf{\Phi}_t|\mathbf{\Theta}_t)$ is JVIM-based shape interpolation via GP mapping. Since $p(\mathbf{\Phi}_t|\mathbf{\Theta}_t)$ is not related to the observation, so the main computational load is the maximization of $p(\mathbf{p}_t, \mathbf{\Theta}_t|\mathbf{I}_t)$. For sequential ATR-Seg, the optimization of $p(\mathbf{p}_t, \mathbf{\Theta}_t|\mathbf{I}_t)$ has two stages: prediction and update. In the first stage (prediction), we use a motion model to predict $p(\mathbf{p}_t, \mathbf{\Theta}_t|\mathbf{I}_{t-1})$ from the previous result $p(\mathbf{p}_{t-1}, \mathbf{\Theta}_{t-1}|\mathbf{I}_{t-1})$ as:

$$p(\mathbf{p}_t, \mathbf{\Theta}_t|\mathbf{I}_{t-1}) = \int\int p(\mathbf{p}_{t-1}, \mathbf{\Theta}_{t-1}|\mathbf{I}_{t-1})p(\mathbf{p}_t|\mathbf{p}_{t-1})p(\mathbf{\Theta}_t|\mathbf{\Theta}_{t-1})d\mathbf{\Theta}_{t-1}d\mathbf{p}_{t-1} \tag{18}$$

where $p(\mathbf{p}_t|\mathbf{p}_{t-1})$ and $p(\mathbf{\Theta}_t|\mathbf{\Theta}_{t-1})$ are used to predict the position and identity/view of a moving target. They are related a motion model that characterizes the target's dynamics and kinematics. In the second stage (update stage), we use the Bayes' rule to compute the posterior as:

$$p(\mathbf{p}_t, \mathbf{\Theta}_t, \mathbf{\Phi}_t|\mathbf{I}_t) = \frac{p(\mathbf{p}_t, \mathbf{\Theta}_t, \mathbf{\Phi}_t|\mathbf{I}_{t-1})p(\mathbf{I}_t|\mathbf{p}_t, \mathbf{\Phi}_t)}{p(\mathbf{I}_t|\mathbf{I}_{t-1})} \tag{19}$$

where $p(\mathbf{p}_t, \mathbf{\Theta}_t, \mathbf{\Phi}_t|\mathbf{I}_{t-1}) = p(\mathbf{p}_t, \mathbf{\Theta}_t|\mathbf{I}_{t-1})p(\mathbf{\Phi}_t|\mathbf{\Theta}_t)$ and we have $p(\mathbf{I}_t|\mathbf{p}_t, \mathbf{\Phi}_t, \mathbf{\Theta}_t) = p(\mathbf{I}_t|\mathbf{p}_t, \mathbf{\Phi}_t)$. Hence, the objective function of the sequential ATR-Seg algorithm can be further rewritten as:

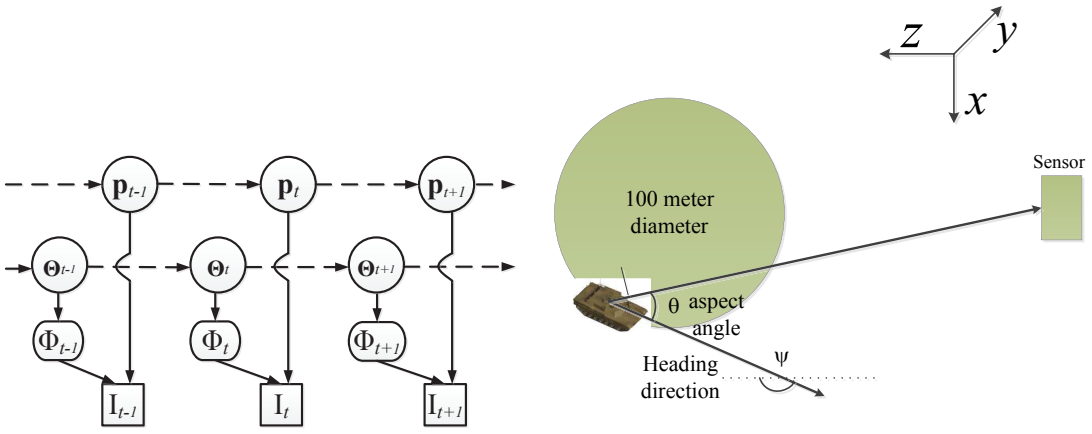
$$\begin{aligned} p(\mathbf{p}_t, \mathbf{\Theta}_t, \mathbf{\Phi}_t|\mathbf{I}_t) \quad \propto \quad & p(\mathbf{I}_t|\mathbf{p}_t, \mathbf{\Phi}_t)\, p(\mathbf{\Phi}_t|\mathbf{\Theta}_t) \iint p(\mathbf{\Theta}_t|\mathbf{\Theta}_{t-1})p(\mathbf{p}_{t-1}, \mathbf{\Theta}_{t-1}, \mathbf{\Phi}_{t-1}|\mathbf{I}_{t-1}) \\ & p(\mathbf{p}_t|\mathbf{p}_{t-1})d\mathbf{\Theta}_{t-1}\, d\mathbf{p}_{t-1} \end{aligned} \tag{20}$$

Due to the nonlinear nature of Equation (20), we resort to a particle filter-based inference framework [29] for sequential optimization, as represented by the graphic model in Figure 7 (left). Thanks to the compact and continuous nature of JVIM, we can draw samples from its latent space for efficient shape interpolation. In the inference process, the state vector is defined as $\mathbf{Z}_t = [\mathbf{p}_t^T, v_t, \psi_t, \alpha_t]^T$, where $\mathbf{p}_t = [p_x^t, p_y^t, p_z^t]^T$ represents the target's 3D position, with the $x-y-z$ axes denoting the horizon (x), elevation (y) and range (z) directions, respectively (as shown in Figure 7 (right)); v_t is the velocity along the heading direction, ψ_t. A 3D-2D camera projection, $W(\mathbf{p})$, is needed to project a 3D position to a 2D position in an image that is assumed to be unchanging for a stationary sensor platform. It is worth noting that we can compute θ_t (the aspect angle) from ψ_t (the heading direction) or *vice versa*. As a matter of fact, the two angles are similar for distant targets when the angle between the lineof sight and the optical axis along the range direction (z) is very small. Because the target is a ground vehicle and to keep it general, a white noise acceleration model is used to represent the dynamics of $\mathbf{Z}_t$, where a simple random walk is applied on the heading direction, ψ_t, to represent arbitrary maneuvering. Moreover, we define the dynamics of α_t (target identity) to be a simple random walk along the identity manifold by which the estimated identity value normally quickly converges to the correct one.

$$\begin{cases} \psi_t &= \psi_{t-1} + \zeta_t^{\psi}, \\ v_t &= v_{t-1} + \zeta_t^{v}, \\ p_x^t &= p_x^{t-1} + v_{t-1}\sin(\psi_{t-1})\Delta t + \zeta_t^{x}, \\ p_y^t &= p_y^{t-1} + \zeta_t^{y}, \\ p_z^t &= p_z^{t-1} + v_{t-1}\cos(\psi_{t-1})\Delta t + \zeta_t^{z}, \\ \alpha^t &= \alpha^{t-1} + \zeta_t^{\alpha}, \end{cases} \tag{21}$$

where Δt is the time interval between two adjacent frames. The process noises associated with the target kinematics, ζ_t^{ψ}, ζ_t^{v}, ζ_t^{x}, ζ_t^{y}, ζ_t^{z}, and ζ_t^{α}, are usually assumed to be a zero-mean Gaussian.

Figure 7. The graphical model representation of ATR-Seg and the 3D camera coordinate. (Reprint from [28] with permission from IEEE).



In a particle filter-based inference algorithm, samples were first drawn according to the dynamics of the state vector and the previous state value, and then, the implicit shape matching defined in Equation (16) was performed to assign a weight for each particle. The mean estimation of weighted

samples produces the solution in the present frame. The pseudo-code for the ATR-Seg algorithm is given in Table 2. Thanks to the unique structure of JVIM, we can capture the continuous and smooth shape evolution during target tracking and recognition, where segmentation $\mathbf{\Phi}_t$ is also archived as a by-product via the shape-aware level set. We expect that the proposed ATR-Seg algorithm has some advantages over other methods that require pre-processing or feature extraction prior to ATR inference [8,26].

Table 2. Pseudo-code for ATR-Seg algorithm.

- Initialization: Initialize the target position, $\mathbf{p}_0$, type α_0, heading direction ψ_0 and speed v_0 according to the ground-truth and get the initial state, $\mathbf{Z}_0$. Draw $\mathbf{Z}_0^j \sim N(Z_0, 1), \forall j \in \{1, \cdots, N_p\}$, N_p is the number of particles.
- For t = 1,$\cdots$, T (number of frames)
  1. For j = 1,$\cdots$, N_p
     1.1 Draw samples $\mathbf{Z}_t^j \sim p(\mathbf{Z}_t^j|\mathbf{Z}_{t-1}^j)$ as in Equation (21).
     1.2 Generate the target shape according to the target state using Equations (9) and (10).
     1.3 Compute weights $w_t^j = p(\mathbf{z}_t|\alpha_t^j, \mathbf{Z}_t^j)$ using Equation (16).
     End
  2. Normalize the weights, such that $\sum_{j=1}^{N_p} w_t^j = 1$.
  3. Compute the mean estimates of the target state, $\hat{\mathbf{Z}}_t = \sum_{j=1}^{N_p} w_t^j \mathbf{Z}_t^j$
  4. Set $\mathbf{Z}_t^j$=resample$(\mathbf{Z}_k^j, w_k^j)$ to increase the effective number of particles [29].
- End

7. Experimental Results

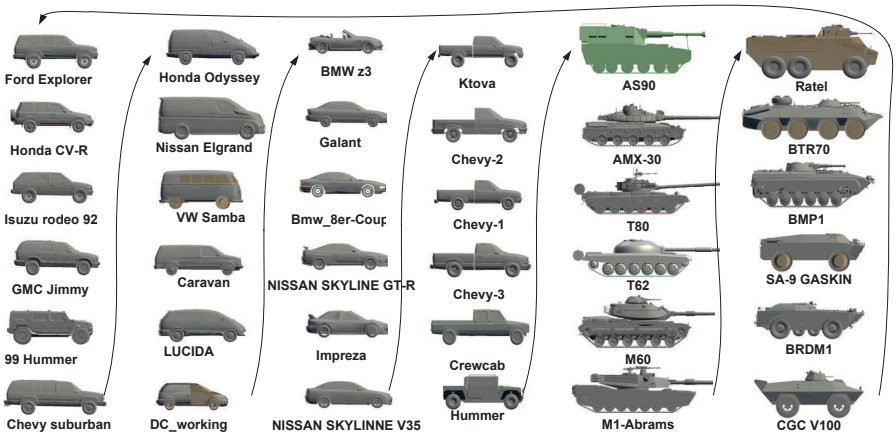
This experimental section provides a detailed evaluation of the ATR-Seg algorithm in six parts. First, we briefly talk about training data collection for JVIM learning with some qualitative results of shape interpolation. Second, we introduce the infrared ATR database used in this work and how different shape models are to be evaluated collectively and fairly. Third, we present the results of the particle filter-based infrared ATR algorithm, where four shape models (JVIM, CVIM, LL-GPLVM, nearest neighbor (NN)) are compared in the case of explicit shape matching. Fourth, we discuss the results of the proposed ATR-Seg algorithm, which involves JVIM-based implicit shape matching and is compared with the algorithms using explicit shape matching (with JVIM and CVIM). Fifth, we discuss the target segmentation results, which are the by-product of the ATR-Seg algorithm. We will also discuss some limitation of ATR-Seg along with some failed cases.

7.1. Training Data Collection

In our work, we considered six target classes as [8], *i.e.*, SUVs, mini-vans, cars, pick-ups, tanks and armored personnel carriers (APCs), each of which has six sub-classes, resulting in a total of 36 targets, as shown in Figure 8. These 36 targets were ordered along the view-independent identity manifold according to a unique topology optimized by the class-constrained shortest-closed-path method proposed

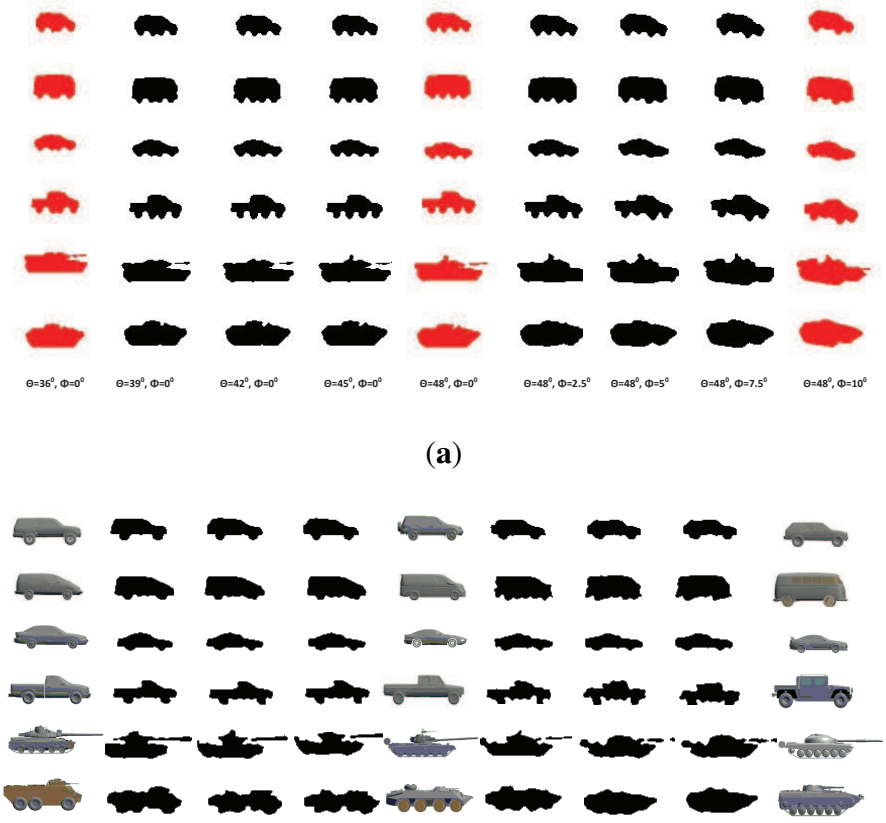
in [8] (before training). We considered aspect and elevation angles in the ranges $0 \leq \theta < 2\pi$ and $0 \leq \phi < \pi/4$, which are digitized in the interval of $\pi/15$ and $\pi/18$ rad, respectively. A total of 150 training viewpoints were used for each target; all training data are generated by their 3D CAD models. In order to reduce the data dimension, the DCT-based shape descriptor proposed in [7] was used here to represent all training shapes for manifold learning. We first detect the contour of a 2D shape (120×80) and then apply the signed distance transform to the contour image, followed by the 2D DCT. Only about 10% DCT coefficients are used to represent a shape, which are sufficient for nearly lossless shape reconstruction. Another advantage of this shape descriptor is that we can do zero-padding prior to inverse DCT to accommodate an arbitrary scaling factor without additional zooming or shrinking operations.

Figure 8. All 36 CAD models used in this work, which are ordered according to the class-constrained shortest-closed-path [8]. (Reprint from [24], with permission from Elsevier.)



JVIM-based shape interpolation is demonstrated in Figure 9, which manifests its capability of handling a variety of target shapes with respect to viewpoint changes for a known target, as well as the generalization to previously unseen target types. In Figure 9a, we pick one target type from each of the six classes. For each target type, we can obtain an identity-specific view manifold from JVIM along which we can interpolate new target shapes of intermediate views (in black) between two training view-points. A smooth shape transition is observed across all interpolated shapes, despite the strong nonlinearity of training shapes. Figure 9b shows the shape interpolation results (in black) along the view-independent identity manifold for the same side view. Although the interpolated shapes are not as smooth as previous ones, most of them are still meaningful, with a mixed nature of two adjacent training target types along the identity manifold. Compared to CVIM in [8], which assumes that the identity and view manifolds are independent, JVIM shows better shape interpolation results by imposing a conditional dependency between the two manifolds and is also more computationally efficient due to local inference. A detailed comparison can be found in [26], where JVIM is found to be advantageous over CVIM and several GPLVM-based shape models, both qualitatively and quantitatively.

Figure 9. Qualitative analysis of JVIM shape interpolation: (**a**) along six identity-specific view manifolds. (**b**) along the view-independent identity manifold between two training target types. (Reprint from [24], with permission from Elsevier.)



(**a**)

(**b**)

7.2. Infrared ATR Database and Shape Models

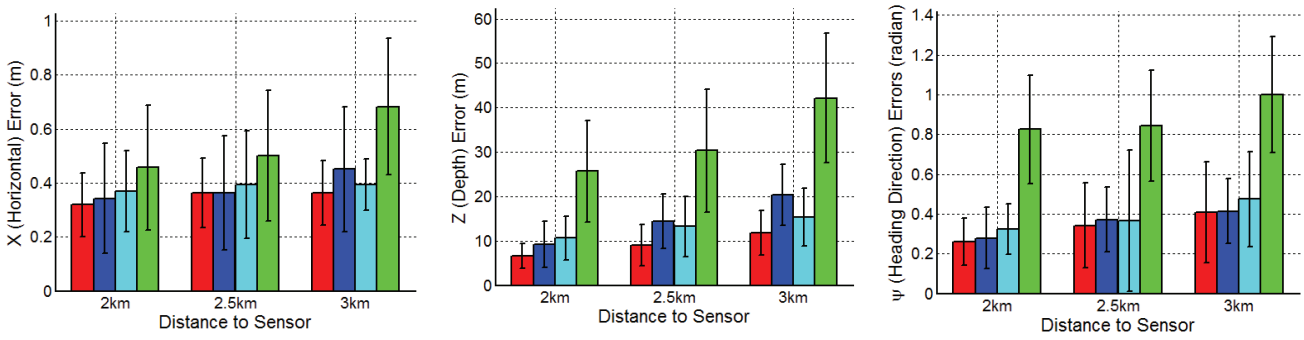
We have obtained a set of mid-wave IR sequences from the SENSIAC ATR database [11], which includes IR imagery of civilian and military ground vehicles maneuvering around a closed-circular path at ranges from 1–3 km. Forty sequences from eight different target types at ranges of 1.0 km, 1.5 km, 2.0 km, 2.5 km and 3 km were selected for this work. For each sequence, tracking was performed on 500 frames. Background subtraction [30] was applied to each frame for clutter rejection, which is needed for two competing algorithms involving explicit shape matching. For each tracking method, the particle filter was initialized with the ground-truth in the first frame. Similar to [8], the process noise of the heading direction ζ_t^{ψ} is assumed to be a non-zero mean Gaussian to accommodate the circular moving trajectory which is necessary due to the ill-posed nature of image-based 3D tracking. This assumption can be relaxed if 3D pose estimation is not needed. Using the metadata provided with the database and a calibrated camera model, we computed the 3D ground-truth of position and aspect angle (in the sensor-centered coordinate system) for each frame. We refer the readers to [26] for more details about the ATR database.

In the following infrared ATR evaluation, we compare JVIM with LL-GPLVM [9] and CVIM [8], as well as the traditional nearest neighbor shape interpolation (NN). Both JVIM and CVIM treat shape factors (view and identity) continuously. To make a fair comparison, we learned a set of target-specific view manifolds by using LL-GPLVM, which involves a hemisphere as the topology constraint for manifold-based shape modeling. Then, we augment a "virtual" circular-shaped identity manifold (similar to that in JVIM and CVIM) for LL-GPLVM, where a NN method is used to "interpolate" arbitrary target types via training ones. Likewise, two "virtual manifolds" are introduced for the NN-based shape model, where we use the nearest neighbor to find the best matched training shapes. Thus, the two shape variables for four shape models can be inferred in a similar continuous way during ATR inference.

7.3. ATR with Explicit Shape Matching

We adopted the particle filter-based ATR algorithm used in [8], where JVIM, CVIM, LL-GPLVM and NN are evaluated in the case of explicit shape matching. In the CVIM-based ATR algorithm, two independent dynamical models are used. In JVIM-based tracking, the dynamic model is a two-stage one, where the first stage is along the view-independent identity manifold, while the second stage along the identity-dependent view manifold. For the LL-GPLVM-based ATR algorithm, one dynamic model is defined on each target-specific view manifold and one on the virtual identity manifold, where NN is used for identity interpolation. For the NN-based ATR algorithm, we employ two dynamic models on two virtual manifolds, like those in CVIM, where shape interpolation is done via NN (*i.e.*, just using the training shapes).

Figure 10. Comparison of the tracking errors of the horizontal position, slant range and heading direction. In each plot, the results for each method averaged over eight target types for each range. From left to right, the plot gives the results for JVIM (first, red), couplet of view and identity manifolds (CVIM) (second, blue), local linear (LL)-Gaussian process latent variable model (GPLVM) (third, cyan) and nearest neighbor (NN) (forth, green). (Reprint from [24], with permission from Elsevier).



The ATR performance of four shape models was evaluated with respect to three figures of merit: (1) $\mathbf{p}_x$ (horizontal) position error (in meters); (2) $\mathbf{p}_z$ (slant range) position error (in meters); and (3) heading direction error ψ (in rads). Quantitative tracking performance results are reported in Figure 10, which give the horizontal, slant range and heading direction tracking errors, respectively,

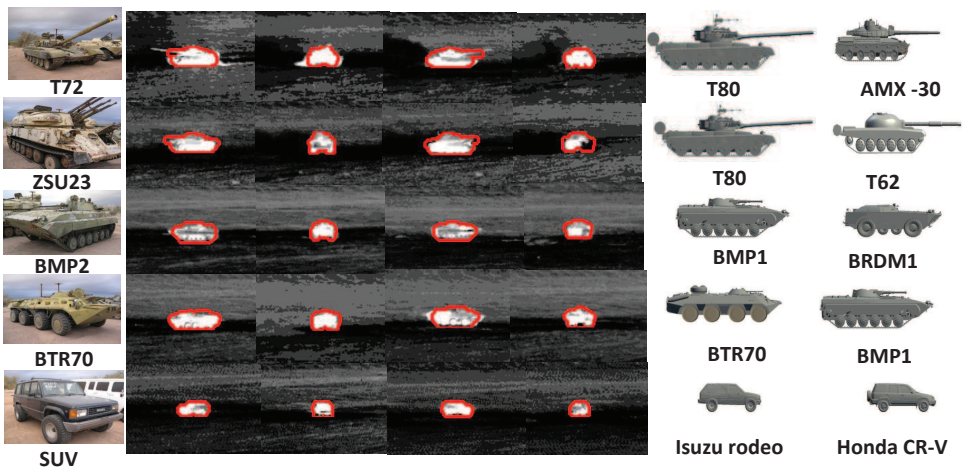
averaged over the eight target types for each range. It is shown that JVIM gains 9%, 10% and 35% improvements over CVIM, LL-GPLVM and NN along the horizontal direction, respectively, 35%, 31% and 72% along the slant range, respectively, and 5%, 13% and 62% along the heading direction, respectively. The results demonstrate that JVIM delivers better tracking performance with respect to all three figures of merit, with the advantage over CVIM, LL-GPLVM and NN being particularly significant for the range estimation.

7.4. ATR-Seg with Implicit Shape Matching

The proposed ATR-Seg algorithm (noted as Method I in the following) was tested against 15 SENSIAC sequences of five targets (SUV, BMP2, BTR70, T72 and ZSU23) under three ranges (1 km, 1.5 km and 2 km). Two more methods, Method II (JVIM with explicit shape matching, [23,26]) and Method III (CVIM [8]), were considered for comparison. All methods share a similar inference algorithm shown in Figure 7. Both Methods II and III involve explicit shape matching, and JVIM was used for both Methods I and II, while CVIM was used for method III. The tracking results are shown in Table 3. Results for tanks were averaged over T72 and ZSU23, and those for APCs averaged over BTR70 and BMP2. It is shown that Method I outperformed Methods II and III by providing lower tracking errors. More importantly, unlike Methods II and III, which require target pre-segmentation, Method I accomplishes target segmentation along with tracking and recognition as a by-product.

During tracking, the target identity is also estimated frame-by-frame by three methods, and the recognition accuracy is calculated as the percentage of frames where the target types were correctly classified in terms of the six target classes. The overall recognition results of three methods are shown in Table 4, where all methods perform well, and Method I (ATR-Seg) still slightly and moderately outperforms Methods II and III, respectively. Especially, when the range is large, e.g., 2 km, the advantage of Method I over Method III is more significant. This is mainly due to the fact that target segmentation is less reliable when the target is small.

Figure 11. ATR-Seg results for five IR sequences. Column 1: truth target types. Columns 2–5: selected IR frames overlayed with the segmentation results. Columns 6–7: the two best matched training targets along the identity manifold. (Reprint from [28], with permission from IEEE.)



The tracking, recognition and segmentation results of Method I (ATR-Seg) against five 1.5-km sequences were shown in Figure 11, where the two best matched target types are presented to show sub-class target recognition. As shown in Figure 11 (the forth tracking result of ZSU23), part of ZSU23 is missing during tracking; the proposed method still can give an accurate segmentation and tracking result. ATR-Seg uses the intensity information from the present frame to build the energy term in Equation (20) that reduces the error accumulation over time and then evaluates how likely a hypothesized shape created by JVIM can segment a valid target at the predicted position. On the other hand, Method III in [8] uses the background subtraction results and involves an explicit shape comparison for evaluation, so the tracking and recognition results highly depend on the pre-segmentation results.

Table 3. Tracking errors for three ATR methods (Method I/Method II/Method III). (Reprint from [28], with permission from IEEE).

Range	Error in	Tank	APC	SUV	Total
1 km	$\mathbf{p}_x\ (m)$	0.22/0.25/0.22	0.22/0.25/0.18	0.16/0.17/0.19	0.21/0.23/**0.20**
	$\mathbf{p}_z\ (m)$	5.06/8.67/7.53	4.19/4.48/5.14	9.03/8.36/10.95	**5.51**/6.93/7.26
	$\psi\ (rad)$	0.13/0.17/0.18	0.15/0.32/0.15	0.11/0.24/0.22	**0.13**/0.24/0.18
1.5 km	$\mathbf{p}_x\ (m)$	0.24/0.19/0.18	0.15/0.21/0.20	0.16/0.56/0.60	**0.27**/0.27/0.27
	$\mathbf{p}_z\ (m)$	4.40/7.20/7.28	4.70/5.88/5.96	$--$NA$--$	**4.55**/6.54/6.26
	$\psi\ (rad)$	0.16/0.22/0.24	0.18/0.53/0.51	0.11/0.32/0.35	**0.20**/0.36/0.37
2 km	$\mathbf{p}_x\ (m)$	0.31/0.27/0.28	0.23/0.19/0.36	0.13/0.17/0.35	0.24/**0.22**/0.32
	$\mathbf{p}_z\ (m)$	8.68/10.6/8.58	8.95/9.28/7.95	5.35/8.09/14.25	**8.19**/9.55/9.46
	$\psi\ (rad)$	0.19/0.38/0.26	0.41/0.18/0.38	0.08/0.41/0.31	**0.26**/0.31/0.32

Table 4. Recognition accuracy (%) for Methods I, II and III. (Reprint from [28], with permission from IEEE).

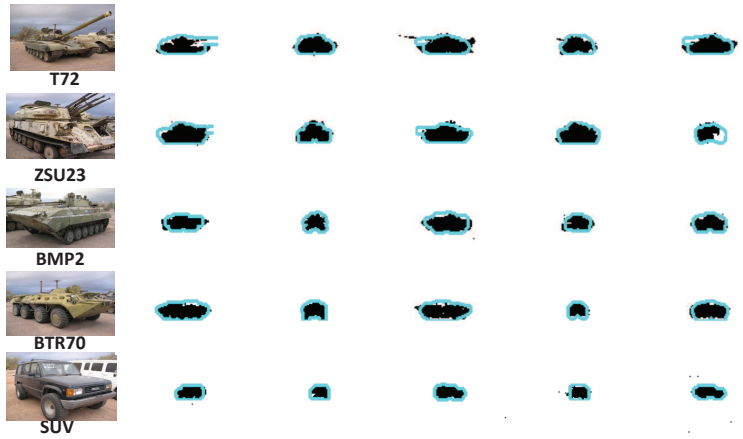
Targets	Tanks	APCs	SUV	Total
1 km	100/96/96	100/100/94	100/100/100	**100**/98/96
1.5 km	98/96/94	99/100/89	100/100/100	**99**/98/93
2 km	98/92/86	100/100/85	100/100/98	**99**/98/88

7.5. ATR-Seg Segmentation Results

We evaluated the segmentation performance of ATR-Seg using the metric of the overlap ratio. The ground-truth segmentation results were generated manually for five randomly selected frames in each of 15 sequences. For a comparison, we also computed the overlap ratios for background subtraction results, which are averaged around 81%. While those of ATR-Seg are averaged around 85%. It is worth noting that the segmentation results of ATR-Seg are essentially constrained by the training shapes created from the CAD models, and the training models may have some shape discrepancy with the

observed targets in the SENSIAC data. Another source of segmentation errors is due to tracking errors. Some segmentation results of five targets at 1.5 km were shown in Figure 12, where we overlaid the ATR-Seg results (contours) over the ground-truth ones. Background subtraction is not easy for a moving platform and is susceptive to the occlusion problem, while ATR-Seg is more flexible and robust to the sensor ego-motion and has great potential for occlusion handling, due to the shape prior involved [6,7].

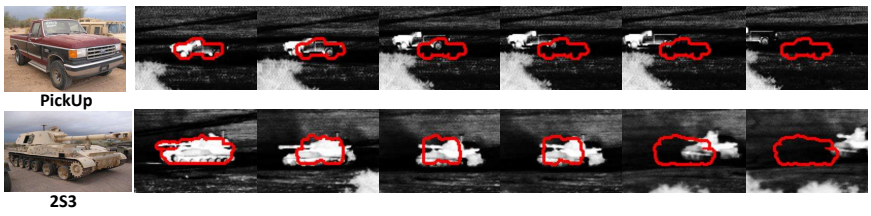
Figure 12. Segmentation results of five targets at the range of 1.5 km.



7.6. Limitation and Discussion

There are two limitations of ATR-Seg due to the unsupervised nature of the level set, where no prior is used for foreground/background pixel intensities, and the mismatching between the training targets and the test ones. Thus, when a major part of a target is occluded or part of a target is similar to the background, the shape-aware set will lose the sensitivity for segmentation evaluation, leading to tracking failure, as shown in Figure 13 (first row), which shows the failed results for the pick-up sequence at 1.5 km. The mismatching and the low-quality data are the main reasons for the tracking failure of 2S3 at a range of 1.5 km (second row in Figure 13). One possible remedy is to incorporate some pixel priors of background and foreground into the level set energy function. However, an online learning scheme may be needed to update the pixel priors that are usually necessary for a long infrared sequence [31]. It is worth emphasizing that the goal of this work is to test the potential of a "model-based" approach that only uses CAD models for training. It is a natural extension to incorporate real infrared data for training that is likely to improve the algorithm robustness and applicability significantly.

Figure 13. Tracking failure for the pick-up and 2S3 sequences at 1.5 km.



8. Conclusion

A new algorithm, called ATR-Seg, is proposed for joint target tracking, recognition and segmentation in infrared imagery, which has three major technical components. First is a novel GPLVM-based shape generative model, the joint view-identity manifold (JVIM), which unifies one view-independent identity manifold and infinite identity-dependent view manifolds jointly in a semantically meaningful latent space. Second is the incorporation of a shape-aware level set energy function that evaluates how likely a valid target can be segmented by a shape synthesized by JVIM. Third, a particle filter-based sequential inference algorithm is developed to jointly accomplish target tracking, recognition and segmentation. Specifically, the level set energy function is used as the likelihood function in the particle filter that performs implicit shape matching, and a general motion model is involved to accommodate a highly maneuvering target. Experimental results on the recent SENSIAC ATR database manifest the advantage of ATR-Seg over two existing methods using explicit shape matching. This work is mainly focused on a shape-based approach. One possible future research issue is to involve other visual cues, such as pixel intensities or textures, to enhance the sensitivity and discriminability of the shape-aware level set energy function, which could mitigate the limitations of the ATR-Seg algorithm.

Author Contributions

This work was supported in part by the U.S. Army Research Laboratory (ARL) and U.S. Army Research Office (ARO) under grant W911NF-04-1-0221 and and the Oklahoma Center for the Advancement of Science and Technology (OCAST) under grant HR12-30.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Yilmaz, A.; Javed, O.; Shah, M. Object tracking: A survey. *ACM Comput. Surv.* **2006**, *38*, 1–45.
2. Fan, X.; Fan, G.; Havilcek, J. Generative Models for Maneuvering Target Tracking. *IEEE Trans. Aerospace Electron. Syst.* **2010**, *46*, 635–655.
3. Srivastava, A. Bayesian filtering for tracking pose and location of rigid targets. *Proc. SPIE* **2000**, *4052*, 160–171.
4. Shaik, J.; Iftekharuddin, K. Automated tracking and classification of infrared images. In Proceedings of International Joint Conference on Neural Networks, Portland, OR, USA, 20–24 July 2003; Volume 2, pp. 1201–1206.
5. Khan, Z.; Gu, I.H. Tracking visual and infrared objects using joint Riemannian manifold appearance and affine shape modeling. In Proceedings of IEEE International Conference on Computer Vision Workshops (ICCVW), Barcelona, Spain, 6–13 November 2011.
6. Prisacariu, V.; Reid, I. Nonlinear shape manifolds as shape priors in level set segmentation and tracking. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 2185–2192.

7. Prisacariu, V.; Reid, I. Shared shape spaces. In Proceedings of IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2587–2594.
8. Venkataraman, V.; Fan, G.; Yu, L.; Zhang, X.; Liu, W.; Havlicek, J.P. Automated Target Tracking and Recognition using Coupled View and Identity Manifolds for Shape Representation. *EURASIP J. Adv. Signal Process.* **2011**, *124*, 1–17.
9. Urtasun, R.; Fleet, D.J.; Geiger, A.; Popović, J.; Darrell, T.J.; Lawrence, N.D. Topologically-constrained latent variable models. In Proceedings of International Conference on Machine learning (ICML), Helsinki, Finland, 5–9 July 2008; pp. 1080–1087.
10. Yao, A.; Gall, J.; Gool, L.; Urtasun, R. Learning Probabilistic Non-Linear Latent Variable Models for Tracking Complex Activities. In Proceedings of Annual Conference on Neural Information Processing Systems (NIPS), Granada, Spain, 12–14 December 2011; pp. 1–9.
11. Military Sensing Information Analysis Center (SENSIAC). Available online: https://www.sensiac.org/ (accessed on 12 December 2012).
12. Yankov, D.; Keogh, E. Manifold Clustering of Shapes. In Proceedings of International Conference on Data Mining, Hong Kong, China, 18–22 December 2006.
13. Roweis, S.T.; Saul, L.K. Nonlinear dimensionality Reduction by Local Linear Embedding. *Science* **2000**, *290*, 2323–2326.
14. Etyngier, P.; Segonne, F.; Keriven, R. Shape Priors using Manifold Learning Techniques. In Proceedings of IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, 14–20 October 2007.
15. Etyngier, P.; Keriven, R.; Segonne, F. Projection onto a Shape Manifold for Image Segmentation with Prior. In Proceedings of IEEE International Conference on Image Processing (ICIP), San Antonio, TX, USA, 16–19 October 2007; Volume 4, pp. 361–364.
16. He, R.; Lei, Z.; Yuan, X.; Li, S. Regularized active shape model for shape alignment. In Proceedings of IEEE International Conference on Automatic Face Gesture Recognition (FG), Amsterdam, the Netherlands, 17–19 September 2008.
17. Lüthi, M.; Albrecht, T.; Vetter, T. Probabilistic Modeling and Visualization of the Flexibility in Morphable Models. In Proceedings of IMA International Conference on Mathematics of Surfaces XIII, York, UK, 7–9 September 2009; Volume 5654, pp. 251–264.
18. Dambreville, S.; Rathi, Y.; Tannenbaum, A. A Framework for Image Segmentation Using Shape Models and Kernel Space Shape Priors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1385–1399.
19. Lawrence, N. Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *J. Mach. Learn. Res.* **2005**, *6*, 1783–1816.
20. Elgammal, A.; Lee, C.S. Separating style and content on a nonlinear manifold. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, 27 June–2 July 2004; Volume 1, pp. 478–485.
21. Lee, C.; Elgammal, A. Modeling View and Posture Manifolds for Tracking. In Proceedings of IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007.

22. Li, X.R.; Jilkov, V. Survey of maneuvering target tracking. Part I. Dynamic models. *IEEE Trans. Aerospace Electron. Syst.* **2003**, *39*, 1333– 1364.
23. Gong, J.; Fan, G.; Yu, L.; Havlicek, J.; Chen, D. Joint view-identity manifold for target tracking and recognition. In Proceedings of 2012 19th IEEE International Conference on Image Processing (ICIP), Orlando, FL, USA, 30 September–3 October 2012; pp. 1357–1360.
24. Gong, J.; Fan, G.; Yu, L.; Havlicek, J.P.; Chen, D.; Fan, N. Joint view-identity manifold for infrared target tracking and recognition. *Comput. Vis. Image Underst.* **2014**, *118*, 211–224.
25. Urtasun, R.; Darrell, T. Sparse probabilistic regression for activity-independent human pose inference. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008.
26. Gong, J.; Fan, G.; Yu, L.; Havlicek, J.P.; Chen, D.; Fan, N. Joint View-Identity Manifold for Infrared Target Tracking and Recognition. *Comput. Vis. Image Underst.* **2014**, *118*, 211–224.
27. Bibby, C.; Reid, I. Robust Real-Time Visual Tracking Using Pixel-Wise Posteriors. In Proceedings of the 10th European Conference on Computer Vision: Part II, Marseille, France, 12–18 October 2008; pp. 831–844.
28. Gong, J.; Fan, G.; Havlicek, J.P.; Fan, N.; Chen, D. Infrared Target Tracking, Recognition and Segmentation using Shape-Aware Level Set. In Proceedings of IEEE International Conference on Image Processing (ICIP), Melbourne, Australia, 15–18 September 2013.
29. Arulampalam, S.; Maskell, S.; Gordon, N.; Clapp, T. A Tutorial on Particle Filters for Online Non-linear/Non-Gaussian Bayesian Tracking. *IEEE Trans. Signal Process.* **2002**, *50*, 174–188.
30. Zivkovic, Z.; van der Heijden, F. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognit. Lett.* **2006**, *27*, 773–780.
31. Venkataraman, V.; Fan, G.; Havlicek, J.; Fan, X.; Zhai, Y.; Yeary, M. Adaptive Kalman Filtering for Histogram-based Appearance Learning in Infrared Imagery. *IEEE Trans. Image Process.* **2012**, *21*, 4622–4635.