JOINT VIEW-IDENTITY MANIFOLD FOR TARGET TRACKING AND RECOGNITION

Jiulu Gong [♯] [‡], *Guoliang Fan* [‡] [*], *Liangjiang Yu* [‡], *Joseph P. Havlicek* [§] *and Derong Chen* [♯].

School of Mechatronical Engineering, Beijing Institute of Technology, China [♯]
School of Electrical and Computer Engineering, Oklahoma State University, USA [‡]
School of Electrical and Computer Engineering, University of Oklahoma, USA [§]

ABSTRACT

A new joint view-identity manifold (JVIM) is proposed for multi-view shape modeling that is applied to automated target tracking and recognition (ATR). This work improves our recent work where the view and identity manifolds are assumed to be independent for multi-view multi-target modeling. A local linear Gaussian process latent variable model (LL-GPLVM) is used to learn a probabilistic JVIM which can capture both inter-class and intra-class variability of 2D target shapes under arbitrary view point jointly in one co-existed latent space. A particle filter-based ATR algorithm is developed to simultaneously infer the view and identity parameters along JVIM so that target tracking and recognition can be achieved jointly in a seamlessly fashion. The experimental results using SENSIAC ATR database demonstrate the advantages of our method both qualitatively and quantitatively compared with existing methods using template matching or separate view and identity manifolds.

1. INTRODUCTION

With the ability to detect, track and recognize both known and unknown targets, automated target tracking and recognition (ATR) is widely used in various military and civilian applications. In vision-based ATR applications, target appearance could change dramatically due to the variations of viewpoint and 3D structure of the target as well as the possibility of unknown target types, which makes ATR a challenging problem in practice. In order to represent the appearance of a 3D rigid object, three major approaches have been used. The first approach suggests a set of representative 2D snapshots [1, 2] captured from multiple viewpoints. Templates [3], histograms [4], edge features [5] etc. are the commonly used non-parametric representation methods for these snapshots. Complex features such as SIFT, HOG, or image patches can be used too. The second approach involves an explicit 3D object model [6] where common representations vary from simple polyhedrons to complex 3D meshes. The third approach uses a manifold learning method to build a low-dimensional nonlinear shape model to capture the shape variability of different objects. For example, in [7] and [8] a nonlinear probabilistic and variational method for adding shape information to level set based object segmentation and tracking was proposed, where GPLVM (Gaussian process latent variable models [9]) is applied to learn a low dimensional latent space and where segmentation and tracing are lined by an image-driven optimization in the latent shape space. However, only one latent factor is considered, either different vehicles under the same view or different poses of the same person. In order to support robust ATR in a 3D scene for both known and unknown targets under an arbitrary view, a general shape model that supports multiple *continuous-valued* factors would be useful and flexible, e.g., identity and view.

In [10, 11] a couplet of identity and view manifolds was applied to ATR where the two manifolds are integrated into a compact target generative model. The main assumption of this method is that the view manifold and identity manifold are independent with the former one pre-designed (a 3D hemisphere) and the latter one learned (a 2D closed-loop), as shown in Fig.1. All target types share the same idealized view manifold. In this paper, we propose a new joint view-identity manifold (JVIM) that captures the coupling effect between the two manifolds for multi-view and multi-class shape modeling. Compared with [10] where the two manifolds are all deterministic, JVIM is probabilistic and involves one co-existed latent space, which is more robust and flexible to handle uncertainty and ambiguity for shape matching. This is demonstrated by the experimental results on the newly released SENSIAC ATR database [12].

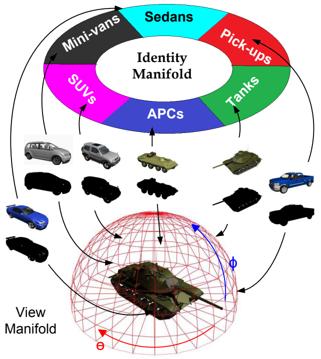


Fig. 1. A couplet view-identity manifolds for shape modeling where a new shape can be interpolated by sampling each manifold [10].

2. PROPOSED METHOD

The key to the JVIM learning is to incorporate certain topology constraints where all latent points can be optimized with respect to both the data terms and the topology prior in a probabilistic manner. In this work, we adopt LL-GPLVM [13] to learn JVIM where the topology priors are inspired from the two manifold structures used in [10]. Moreover, we make two efforts to reduce the complexity of LL-GPLVM learning. One is to use a DCT-based shape descriptor recently proposed in [7] to reduce the dimensionality of input shapes, and the other is to invoke a local point approximation-based learning method [14]. To facilitate the inference process, JVIM can be further represented into one *view-independent identity manifold* and an *identity-dependent view manifold* given an identity hypothesis, as shown by Fig. 2. Then a similar particle filtering-based ATR inference method in [10] can be used to recognize a target (at both class and sub-class levels) and to track the 3D position/pose simultaneously.

[*] Corresponding author,

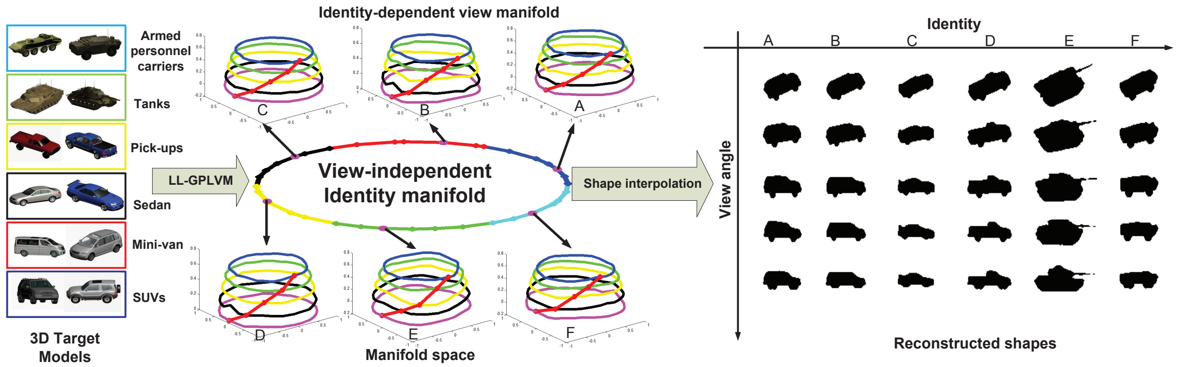


Fig. 2. Illustration of the generative model for shape interpolation along the manifold.

2.1. Target Representation

All shapes in this work were initially represented implicitly by using the signed distance function ($120 \times 80 = 9600$), and then we apply the 2D DCT-based shape descriptor [7] to reduce the dimensionality of training data, making the GLPVM learning more tractable. As shown in Fig.3, the DCT quantization with $30 \times 30 = 900$ coefficients preserved can reconstruct a shape with a reasonable quality.

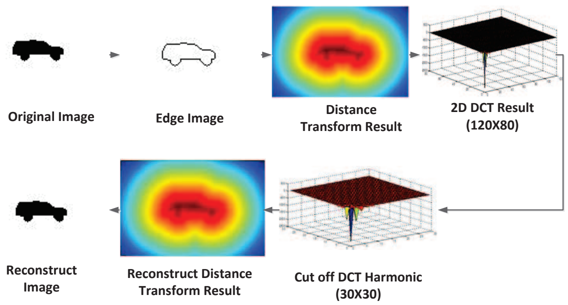

Original Image
Edge Image
Distance Transform Result
2D DCT Result (120X80)
Reconstruct Image
Reconstruct Distance Transform Result
Cut off DCT Harmonic (30X30)

Fig. 3. DCT-based shape representation where only less 10% DCT coefficients are used for shape reconstruction.

2.2. GPLVM and LL-GPLVM

GPLVM represents a set of high-dimensional (HD) data ,$\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}^T$ and $\mathbf{y}_i \in R^d$, in a low-dimensional (LD) latent space, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}^T$ and $\mathbf{x}_i \in R^q$, and a Gaussian process is used as the nonlinear mapping from the latent space to the data space where a covariance function is $(\mathbf{K_Y})_{ij} = k_\mathbf{Y}(\mathbf{x}_i, \mathbf{x}_j)$ is involved. The learning of GPLVM seeks to maximize the likelihood of the data given the latent positions,

$$p(\mathbf{Y}|\mathbf{X}, \beta) = \frac{1}{Z}\exp(-\frac{1}{2}tr(\mathbf{K}_\mathbf{Y}^{-1}\mathbf{Y}\mathbf{Y}^T)), \quad (1)$$

where Z is a normalization factor, $\mathbf{K}_Y$ is known as the kernel matrix, and β denotes the kernel hyper-parameters. In order to reflect the natural (or intrinsic) topology of the data or to encourage a given manifold topology, the local linear GPLVM (LL-GPLVM) was proposed in [13] for human motion modeling. LL-GPLVM introduces topological constraints based on a neighborhood structure learned from local linear embedding (LLE) [15]. The objective function for LL-GPLVM is

$$\begin{aligned} Ls &= \log p(\mathbf{Y}|\mathbf{X}, \beta)p(\beta)p(\mathbf{X}|\mathbf{w}) \\ &= \frac{d}{2}\ln|\mathbf{K_Y}| + \frac{1}{2}tr(\mathbf{K_Y}^{-1}\mathbf{Y}\mathbf{Y}^T) + \sum_i \beta_i \\ &\quad + \frac{1}{\sigma^2}\sum_{i=1}^N \|\mathbf{x}_i - \sum_{j=1}^N \omega_{ij}\mathbf{x}_j\|^2 + C, \end{aligned} \quad (2)$$

where C is a constant, ω_{ij} is the weights that best reconstruct each data point $\mathbf{x}_i$ from its neighbors $\mathbf{x}_j$ by minimizing $\Phi(w) = \sum_{i=1}^N \|\mathbf{x}_i - \sum_{j=1}^N \omega_{ij}\mathbf{x}_j\|^2$, σ^2 is used to adjust the strength of topology constrained. After learning, given a LD point $\mathbf{x}_{test}$, its corresponding HD distribution is defined by

$$p(\mathbf{y}|\mathbf{x}_{test}, \mathbf{Y}, \mathbf{X}, \beta) = N(\mathbf{y}|\mu, \upsilon), \quad (3)$$

where $\mu = \mathbf{k}_\mathbf{x}^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{Y}$, $\upsilon = \mathbf{K}_{\mathbf{xx}} - \mathbf{k}_\mathbf{x}^T(\mathbf{K} + \sigma^2\mathbf{I})\mathbf{k}_\mathbf{x}$ and $\mathbf{k}_\mathbf{x} = k(\mathbf{x}_{test}, \mathbf{x})$. In this work, μ is used for shape reconstruction, and υ indicates the reconstruction confidence.

2.3. JVIM Learning by LL-GPLVM

One important feature of JVIM is that its latent space is semantically meaningful, and it is ensured through the topology constraints to be involved in LL-GPLVM. As discussed in [10, 11] a 2D hemisphere-shaped view manifold shared by all target types is used to deal with view variations by spanning all possible view angles for ground vehicles and a 1D circular-shaped identity manifold is used to captures both inter-class and intra-class shape variability. Let $\mathbf{y}_m^k \in R^d$ to be the $d-$dimensional (d=900) data of target k under view m, and $\mathbf{x}_m^k = [\theta_m^k, \phi_m^k, \alpha^k]$, $0 \leq \theta_m^k \leq 2\pi$, $0 \leq \alpha^k \leq \pi$, $0 \leq \theta_m^k \leq 2\pi$ denote the corresponding point in the LD space. The first two parameters θ_m^k and ϕ_m^k represent the azimuth and elevation angles of view m respectively, and the last one α^k is the identity term. We can separate the LD space into two parts $\mathbf{x} = [\mathbf{x}_{view}, \mathbf{x}_{id}]$ and $\mathbf{x}_{view}^k = [\theta_m^k, \phi_m^k]$, $\mathbf{x}_{id} = \alpha$, the first part can be treated as an *identity-dependent view manifold*, and the second part is an *view-independent identity manifold.* As shown in Fig. 4 for each identity-dependent view manifold a hemisphere structure is used to initialize JVIM. It's worth noting that both view and identity factors are jointly optimized during the LL-GPLVM learning process.

An important issue in LL-GPLVM is the selection of neighborhood for each data point that is preferred to be done in the LD latent space. In JVIM learning, we choose the neighborhood for each training data point in the latent space according to the semantic structure of JVIM. As in Fig. 4, for a training point (red point) we have two

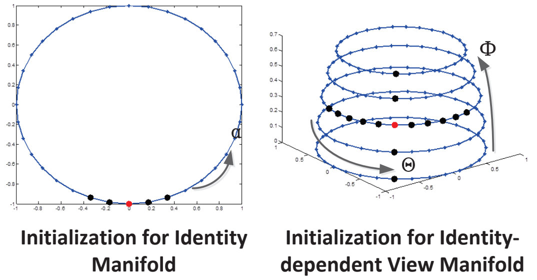

α
Φ
Θ
Initialization for Identity Manifold
Initialization for Identity-dependent View Manifold

Fig. 4. Initialization structure and neighborhood structure for JVIM.

steps to choose neighbors. The first step (left) is to choose some neighbors (4 black points) along the α (identity) direction without changing the view parameters, and the second step (right) is to select some neighbors (14 black ones) along θ and ϕ directions under the same identity parameter. All neighbor points (18 in this work) will be used to calculate the LLE weights ω_{ij} in (2). The weights can be computed by solving, $\forall j \in \eta_i$, the following system,

$$\sum_k C_{kj}\omega_{ij} = 1, \tag{4}$$

where $C_{kj} = (\mathbf{x}_i - \mathbf{x}_k)^T(\mathbf{x}_i - \mathbf{x}_j)$ if $j, k \in \eta_i$, and 0 otherwise. Then scaling them to satisfy $\sum_j \omega_{ij} = 1$. C_{kj} is computed using the neighborhood shown in Fig. 4 as follow,

$$\left\{\begin{array}{lcl} C_{kj} & = & C_{kj}^{x1} + C_{kj}^{x2} + C_{kj}^{x3} + C_{kj}^{x4} + C_{kj}^{x5} \\ C_{kj}^{x1} & = & (\cos(\phi_i)\cos(\theta_i) - \cos(\phi_k)\cos(\theta_k)) \\ & & (\cos(\phi_i)\cos(\theta_i) - \cos(\phi_j)\cos(\theta_j)) \\ C_{kj}^{x2} & = & (\cos(\phi_i)\sin(\theta_i) - \cos(\phi_k)\sin(\theta_k)) \\ & & (\cos(\phi_i)\sin(\theta_i) - \cos(\phi_j)\sin(\theta_j)) \\ C_{kj}^{x3} & = & (\sin(\phi_i) - \sin(\phi_k))(\sin(\phi_i) - \sin(\phi_j)) \\ C_{kj}^{x4} & = & (\cos(\alpha_i) - \cos(\alpha_k))(\cos(\alpha_i) - \cos(\alpha_j)) \\ C_{kj}^{x5} & = & (\sin(\alpha_i) - \sin(\alpha_k))(\sin(\alpha_i) - \sin(\alpha_j)), \end{array}\right. \tag{5}$$

where θ, ϕ and α are the azimuth angle, elevation angle in the identity-dependent view manifold and identity parameter in the identity manifold in Fig. 4.

2.4. Efficient JVIM Learning

A computational bottleneck of GPLVM learning is the gradient-based iterative optimization method where the inverse of $\mathbf{K}_Y$ in (1) is involved and could be computationally prohibitive for a large training data set. There are two approaches to address this issue. One is to use an active set for sparse learning where the active set is optimized during each iteration [16], and the other is to use a local approximation to define $\mathbf{K}_Y$ for each point via its neighbors [14]. In this paper, we use the later one that is more efficient and simpler. More importantly, the topology-awareness of JVIM facilitates the neighborhood selection for the local approximation to $\mathbf{K}_Y$, and makes JVIM amenable to this approach. Similar to the topological constraint shown in Fig. 4, we expand the neighborhood definition for each data point by including more neighbors to compute the local $\mathbf{K}_Y$ matrix for efficient JVIM learning as well as JVIM-based shape interpolation.

3. EXPERIMENTAL RESULT

We used the newly released SENSIAC ATR database to test the proposed JVIM. 24 night-time midwave infrared (MWIR) sequences (8 vehicles at 3 ranges, 2km, 2.5km and 3km) were used. The background substraction was applied to obtain the initial target segmentation results. Three methods were tested, including the proposed one (Method-I), the one in [10] (Method-II), and the traditional template-based method without shape interpolation (Method-III).

3.1. Model learning

We used the same 36 3D CAD models for JVIM training as in [10], six for each of the six target types (APCs, tanks, pick-ups, cars, minivans, SUVs), as shown in Fig. 5, where 36 target types are ordered according to a unique topology obtained by finding a shortest-closed-path [10]. For each 3D model, we generate a set of training data. In this work, 12° and 10° intervals along the azimuth and elevation angles were used and 150 training viewpoints were obtained for each target. The learned JVIM can be examined in terms of its capability of shape interpolation alone the joint manifold. Fig.6 shows the interpolation result, which shows semantically meaningful interpolated shapes, making it possible to handle not only new viewpoints for a known target but also arbitrary viewpoints for unknown targets.

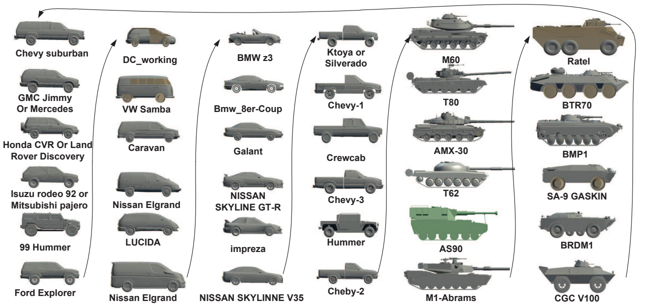
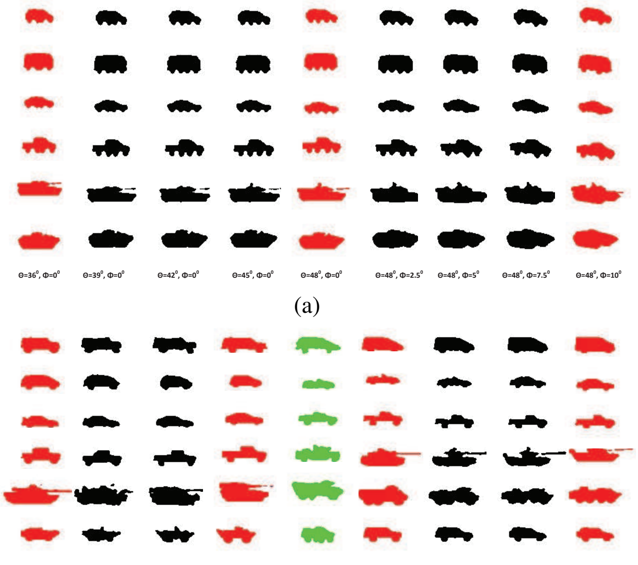


Fig. 5. All 36 3D CAD models used for learning.


$\Theta=36^0$, $\Phi=0^0$
$\Theta=39^0$, $\Phi=0^0$
$\Theta=42^0$, $\Phi=0^0$
$\Theta=45^0$, $\Phi=0^0$
$\Theta=48^0$, $\Phi=0^0$
$\Theta=48^0$, $\Phi=2.5^0$
$\Theta=48^0$, $\Phi=5^0$
$\Theta=48^0$, $\Phi=7.5^0$
$\Theta=48^0$, $\Phi=10^0$

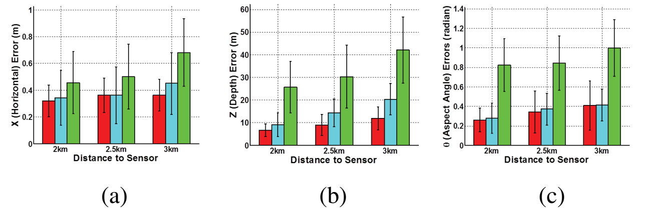
(a)

(b)

Fig. 6. JVIM-based shape interpolation along the view manifold (a) and along the identity manifold (side-view only) (b).

3.2. Target Tracking

The tracking and recognition efficiency was validated against SENSIAC ATR database, which includes a rich set of meta data for each frame of every sequence and can be used to calculate the ground truth data. The performance of our method was evaluated based on the error in the estimated 3D position and aspect orientation. For each sequence we chose approximately 1000 frames for testing, and only the other frame was used for tracking. A total of about 500 frames were used. We computed the error in estimated target position along the x (horizon) and z (distance from the camera), as well as the aspect orientation. We initialized the tracking by the ground truth data for the first frame, and the overall tracking result is shown in Fig.7, the result is the average over 8 targets of the same range. Our result outperforms the one proposed in [10]. We also visualized some detailed tracking results of our proposed method against eight 2000m sequences in Fig. 8, where we have overlaid the interpolated shapes on the target according to the estimated 3D position and aspect angle as well as the estimated target type.



(a) (b) (c)

Fig. 7. Overall 3D tracking performance of Methods-I, II and III (from the first to the third bars). (a) Horizontal errors (m). (b) Range errors (m). (c) Aspect angle errors (rad).

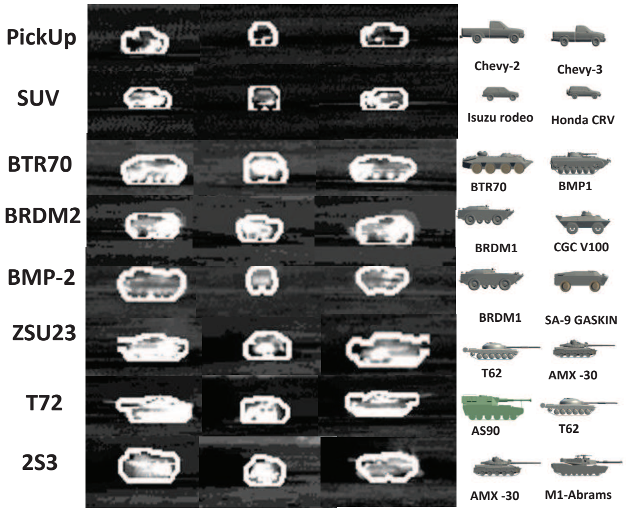


Fig. 8. Target tracking results showing actual SENSIAC IR frames overlaid with interpolated shapes produced.

3.3. Target Recognition

Since the target type is estimated frame by frame during tracking, we define the overall recognition accuracy as the percentage of frames where the target is correctly classified in terms of the six classes. A circle like identity manifold is used in this paper so the identity factor becomes an angular one $\alpha \in [0, 2\pi)$ for each target. Therefore, tanks, APCs, SUVs, pick-ups, minivans and cars, each will be represented by an angular section along the identity manifold. The overall recognition results of the three methods for 24 sequences are shown in Table 1, where the accuracy of Tanks is averaged over the T72, ZSU23, and 2S3 target types and that of the APCs is averaged over those of BTR70, BMP2, and BRDM2 target types. Our recognition result is significantly better than Method-II and Method-III. We also present the two best matched training targets on the right side in Fig. 8 to demonstrate the sub-class target recognition capability, where the best matches often include the correct sub-class.

Table 1. Overall recognition accuracies (%) of four methods (Method-I/Method-II/Method-III) against 24 SENSIAC sequences.

Targets	Tanks	APCs	SUV	Pick-up
2000m	93/86/81	100/85/80	100/98/95	100/97/95
2500m	90/78/69	92/76/70	94/92/86	100/90/86
3000m	81/70/60	89/72/65	100/86/79	100/82/77

4. CONCLUSION

A new joint view and identity manifold (JVIM) has been proposed for multi-view and multi-class shape modeling in the same latent space. The LL-GPLVM algorithm used to learn JVIM where the local approximation method is used to speed up the learning process and shape interpolation. The experiment results on IR data show the advantage of our method over the recent one in [10] where two manifolds are assumed to be independent. Our future work will focus on the integration of segmentation into the ATR process, making the proposed JVIM more useful for real-word applications.

5. REFERENCES

[1] T. Poggio and S. Edelman, “A network that learns to recognize three-dimensional objects,” *Nature*, vol. 343, pp. 263–266, 1990.

[2] S. Ullman and R. Basri, “Recognition by linear combinations of models,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, pp. 992–1006, 1991.

[3] M. Xue, S. K. Zhou, and H. Wu, “Integrated detection, tracking and recognition for IR video-based vehicle classification,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing.*, 2006.

[4] V. Venkataraman, G. Fan, and X. Fan, “Target tracking with online feature selection in FLIR imagery,” in *Proc. IEEE Workshop on Object Tracking and Classification Beyond Visible Spectrum (OTCBVS07) in conjunction with CVPR07*, 2007.

[5] J.S. Shaik and K.M. Iftekharuddin, “Automated tracking and classification of infrared images,” in *Proc. International Joint Conference on Neural Networks*, 2003.

[6] S. Ullman, “An approach to object recognition: Aligning pictorial descriptions,” *Cognition*, vol. 32, pp. 193–254, 1989.

[7] V. Prisacariu and I. Reid, “Shared shape spaces,” in *Proceeding of the International Conference On Computer Vision*, 2011.

[8] V. Prisacariu and I. Reid, “Nonlinear shape manifolds as shape priors in level set segmentation and tracking,” in *Proceeding of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[9] N Lawrence, “Probablistic non-linear principle component analysis with gaussian process latent variable model,” *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[10] V. Venkataraman, G. Fan, L. Yu, X. Zhang, W. Liu, and J. P. Havlicek, “Automated target tracking and recognition using coupled view and identity manifolds for shape representation,” *Advances in Signal Processing*, 2011.

[11] V. Venkataraman, G. Fan, L. Yu, X. Zhang, W. Liu, and J. P. Havlicek, “Joint target tracking and recognition using view and identity manifolds,” in *Proc. IEEE Workshop on Object Tracking and Classification Beyond Visible Spectrum in conjunction with CVPR2011*, 2011.

[12] “Military Sensing Information Analysis Center (SENSIAC),” 2008, https://www.sensiac.org/.

[13] R. Urtasun, D. J. Fleet, A. Geiger, J. Popovic, T. J. Darrell, and N. D. Lawrence, “Topologically-constrained latent variable models,” in *Proceeding of the International Conference On Machine Learning*, 2008.

[14] A. Yao, J. Gall, L. Gool, and R Urtasun, “Learning probabilistic non-linear latent variable models for tracking complex activities,” in *Proceeding of Twenty-Fifth Annual Conference on Neural Information Processing Systems*, 2011.

[15] S. t. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by local linear embedding,” *Science*, vol. 290, pp. 2323–2326, 2000.

[16] Q. Joaquin and C. E. Rasmussen, “A unifying view of sparse approximate gaussian process regression,” *Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.