Joint Target Tracking and Recognition using View and Identity Manifolds

Vijay Venkataraman †, Guoliang Fan †, Liangjiang Yu †, Xin Zhang †, Weiguang Liu ‡, and Joseph P. Havlicek §
School of Electrical and Computer Engineering, Oklahoma State University †
College of Computer Science, Zhongyuan University of Technology, China ‡
School of Electrical and Computer Engineering, University of Oklahoma §

Abstract

We propose a new concept of identity manifold for automated target tracking and recognition (ATR) that captures both inter-class (e.g., between tanks and armored cars) and intra-class (e.g., between different tanks) variability of target appearances (e.g., shapes). A hemisphere-shaped view manifold is also involved for mutli-view target modeling. Combining the two continuous-valued manifolds via non-linear tensor decomposition gives rise to a new generative model that can be learned from a small training set. This model can not only deal with arbitrary view/pose variations by tracking along the view manifold, but also interpolate the appearance of an unknown target along the identity manifold. The proposed model is examined based on the recently released SENSIAC ATR database, and the experimental results confirm the usefulness of this generative model.

1. Introduction

The main challenge in vision-based automated target tracking and recognition (ATR) is the variation of target appearances due to different viewpoints and various body structures. Usually these two factors, as especially the second one, are considered as discrete variables [17, 16]. In this work, we want to account for both factors in a continuous manner by using view and identity manifolds that can be learnt from a small training set. The use of the two manifolds facilitates the ATR inference process, and allows us to meaningfully synthesize new target appearances to deal with unknown targets under unseen viewpoints. According to [23], object recognition research can be roughly grouped into three categories, viz., single-view 2D models, single instance 3D models, and multi-view models. The first group of methods focus on object detection rather than pose estimation by modeling the appearances of multiple objects in a single, discrete or limited range of views [6, 31] without relating features across multiple views. Those of the second group estimate the pose/view by matching local features under rigid transformations [15, 21], making extensions to other object classes difficult. Those of the third group aim to build a coherent object model by relating descriptive features over multiple views [9, 13, 12, 24]. Our research belongs the third group.

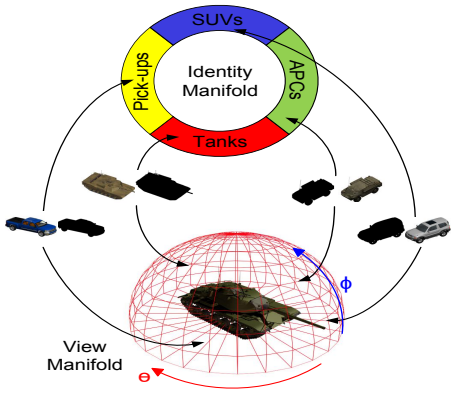


Figure 1. Coupled view-identity manifolds for shape-based multi-view target modeling. We decompose the shape variability in the training set into the two factors, identity and view, both of which can be mapped to a low dimensional manifold. Then by choosing a point on each manifold, a new shape can be interpolated.

In particular, we propose a shape-based generative target model that is controlled by the continuous view and identity variables and involves two manifolds for multi-view target modeling, as shown in Fig.1. The first is the 1-D *identity manifold* that is learnt to represent both inter-class and intra-class shape variability among all training targets. The second is a 2D hemisphere-shaped view manifold to cope with arbitrary view variations. A non-linear tensor decomposition technique is used to integrate the two manifolds into a compact generative model that can be incorporated in a particle filter for ATR tasks. Our research is based on the new ATR database released by the Military Sensing Information Analysis Center (SENSIAC) [1] that contains a rich set of infrared imagery of various military and civilian vehicles. Specifically, we choose four target classes in this work according to the SENSIAC dataset, i.e., *tanks*, *armored personnel carriers (APCs)*, *SUVs*, and *Pick-ups*, each of which includes several sub-classes. The experimental results demonstrate the advantages of the proposed method over the traditional template-based ones.

2. Related Work

There are two theories on object representation. One suggests a set of representative 2D snapshots [20, 29] and the other involves a 3D object model [28]. In the first theory, unknown views can be interpolated from the given ones, while in the second one, the 3D model is used to match the 2D observation via 3D-to-2D projection. Accordingly, most object recognition methods can be categorized into two groups: those involving 2D multi-view images [19, 24, 23, 26, 9, 8] and those supported by explicit 3D models [14, 11, 22, 27]. Some make use of both the 3D shape and 2D appearances [13]. A variety of 2D features (e.g., silhouettes, edges, HOG, SIFT) or 3D models (e.g., meshes, polyheadrons) were used in these methods. The psychophysical evidence [3] motivates us to use 2D view-based silhouettes for multi-view object representation.

One interesting issue of view-independent object recognition is to determine the low dimensional (LD) embedding of the latent factors (e.g., view and identity) from the high-dimensional (HD) observations. This LD embedding allows us to synthesize object appearances for unknown views or to accommodate some intra-class appearance variability. One early work in [18] applied PCA to find two separate eigenspaces (one for object identities and one for a specific object under different poses) for joint identity and pose estimation. The bilinear models [25] and multilinear analysis [30] have provided more systematic multi-factor representation by decomposing HD observations into several independent factors. In [7], the view variable is related with the appearance through shape sub-manifolds which have to be learned for each object class. Our work was inspired by the non-linear tensor decomposition proposed in [10] that involves a radial basis function (RBF)-based non-linear mapping prior to multi-factor tensor analysis. Specifically, we can decouple the view and identity variables by imposing an *identity-independent view manifold* and learning a *view-independent identity manifold*.

Along another line of thinking, some methods can synthesize novel 3D shapes from a set of 3D objects belonging to the same class. For example, in [27] a 3D object is represented by mesh vertices that are matched to salient feature points in the observed image. The main challenge is to determine correspondences between the model and observation. A correspondence-free method was presented in [22], where multiple 3D objects were used to develop a generic 3D shape prior and PCA was applied to generate novel 3D shapes. Then the shape and pose parameters can be optimized jointly by maximizing the degree of match between the 2D projection of the novel 3D model given the segmentation boundary of the unknown object. In contrast, our generative model is controlled by two independent variables constrained on their own LD manifolds, making the inference process very efficient and flexible.

3. Generative Target Models

Although our approach can be applied to different object tracking/recognition problems, we study infrared ATR in this work where the silhouette-based shape feature is used to learn the new generative target model.

3.1. Identity manifold

The identity manifold plays a central role in our work that is intended to capture both inter-class and intra-class shape variability among training targets. More importantly, it is unlike other methods in terms of the way that the identity is handled: the continuous nature of the proposed identity manifold makes it possible to interpolate an unknown target type based on limited training data. There are two important questions to be addressed in order to learn such a manifold with the desired capability. The first one is where or in which space this identity manifold should be learned. Ideally, it should be learned in a LD latent space with only the identity factor rather than in the HD observation space where the view and identity factors are coupled together. The second one is how to learn a *semantically valid* identity manifold that supports meaningful shape interpolation for an unknown target. In other words, what kind of constraint should be imposed on the identity manifold to ensure that the interpolated shapes represent real-world targets.

We defer the discussion of the first issue to Section 3.3, while here we focus on the second one that involves the determination of an appropriate manifold topology, including the dimension and the topological relationship among all training targets. In particular, we suggest a *1D closed-loop structure* in this work. There are a few considerations for this seemingly arbitrary but actually practical choice. First, the learning of a higher-dimensional manifold requires a large set of training targets that may not be needed for a specific ATR application which may have a small set of targets-of-interest. Second, this identity manifold is assumed to be "closed" rather than "open," because all targets of interest here are man-made ground vehicles which share some degree of similarity with extreme disparity unlikely. Third, the 1D closed structure would greatly facilitate the inference process for online ATR tasks. As a result, the manifold topology is reduced to a specific *ordering relationship* of training targets along the 1D closed identity manifold. Ideally, we want targets of the same class or of similar shapes to stay close compared with dissimilar ones. Thus we introduce a *class-constrained shortest-closed-path* method to find a unique ordering relationship across all targets along the identity manifold. This method requires a view-independent *distance* or *dissimilarity* measure between two targets. For example, we could use the shape dissimilarity between two 3D target models that can be approximated by the accumulated mean square errors of multi-view silhouettes (after the distance transform [4]).

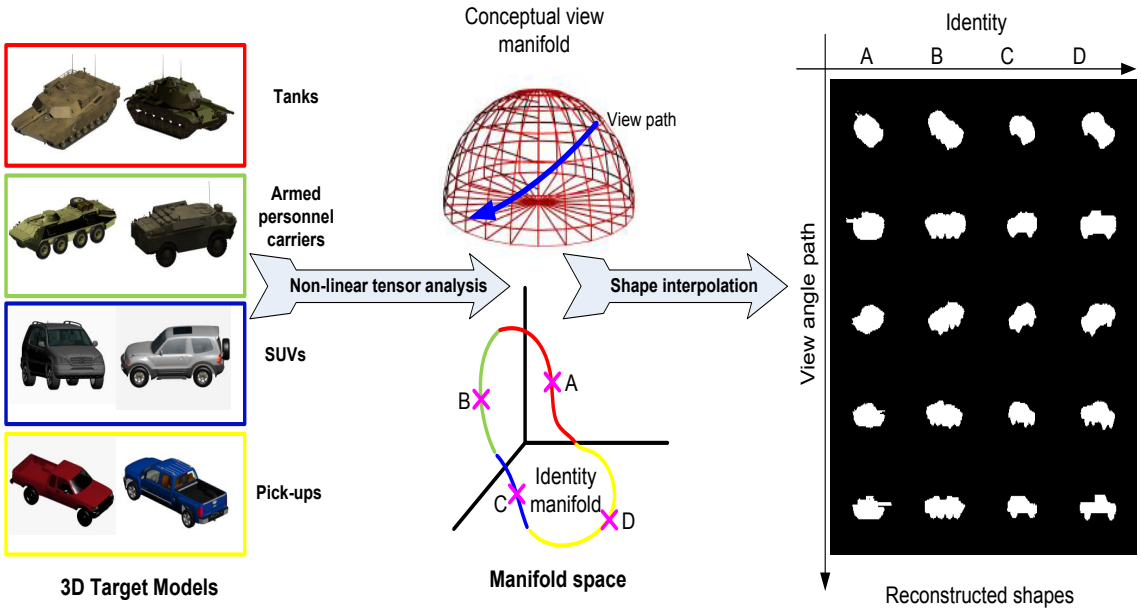


Figure 2. Illustration of the generative model for view and identity based shape appearance synthesis. Reconstruction results of the shape are shown for the blue path traversed along the view manifold and for four different points on the identity manifold that do not correspond to any of the training data. In each case the shape reconstructed has strong characteristics of the view and vehicle class.

Given N individual targets, suppose that we can find a set of N view-independent identity vectors in a LD latent space $\boldsymbol{i}^k, k \in \{1, \cdots, N\}$ along with the associated class labels $\boldsymbol{L}_k$. Let $\boldsymbol{y}_m^k$ denote the vectorized silhouette of object k under view m. The similarity between targets u and v, represented by $\boldsymbol{i}^u$ and $\boldsymbol{i}^v$ respectively, over all M training views is given by

$$D(\boldsymbol{i}^u, \boldsymbol{i}^v) = \sum_{m=1}^{M} ||\boldsymbol{y}_m^u - \boldsymbol{y}_m^v|| + \alpha \cdot \epsilon(\boldsymbol{L}_u, \boldsymbol{L}_v), \quad (1)$$

and where

$$\epsilon(\boldsymbol{L}_u, \boldsymbol{L}_v) = \begin{cases} 0 & \text{if } \boldsymbol{L}_u = \boldsymbol{L}_v, \\ 1 & \text{otherwise,} \end{cases} \quad (2)$$

where $||.||$ represents the Euclidean distance and α is a constant. The function $\epsilon(\boldsymbol{L}_u, \boldsymbol{L}_v)$ is a penalty term that ensures targets belonging to the same class are grouped together along the identity manifold. Let the manifold topology be represented by $\mathbf{T} = [t_1\ t_2\ \cdots\ t_{N+1}]$ where $t_i \in [1, N]$, $t_i \neq t_j$ for $i \neq j$ and $t_1 = t_{N+1}$. The class-constrained shortest-closed-path can be written as

$$\mathbf{T}^* = \arg\min_{\mathbf{T}} \sum_{i=1}^{N} D(\boldsymbol{i}^{t_i}, \boldsymbol{i}^{t_{i+1}}). \quad (3)$$

This manifold topology tends to group those targets of similar 3D shapes and/or the same class together, enforcing the best local *semantic smoothness* along the identity manifold to be learned, which is essential for valid shape interpolation of an unknown target type.

3.2. Conceptual view manifold

Additionally, we also need to specify a view manifold to accommodate the view-induced shape variability. A common way is to use some non-linear dimensionality reduction techniques, such as LLE or Laplacian eigenmaps, to find the LD view manifold from the HD observations for a given target [4]. There are two main limitations. One is that they are identity-dependent, and multiple view manifolds are involved which may have to be aligned together in the same latent space for general multi-view target modeling. The other is that they are normally constrained by a 1D structure that may not reflect all possible object poses in the real-world. In this paper, the view manifold is designed to be a hemisphere that embraces almost all possible viewing angles around a ground vehicle as shown in Fig. 1 and is characterized by two parameters: the azimuth angle θ and the elevation angle ϕ. Such a conceptual manifold helps us avoid the issue with learning and aligning multiple view manifolds for different targets. At the same time, it provides a unified and intuitive representation of the view space and supports efficient dynamic view estimation.

3.3. Non-linear Tensor Decomposition

We extend the non-linear tensor decomposition in [10] to develop the proposed generative target model. This involves learning a non-linear mapping function from the HD observations to the unified view manifold and then factoring out identity vectors, giving a view-independent space for identity representation (the first question raised in Section 3.1).

Let $\boldsymbol{y}_m^k \in \mathcal{R}^d$ that is the d-dimensional observation of target k under view m) and $\boldsymbol{x}_m$ denote the LD point of view m on the view manifold. For target k, we can learn a non-linear mapping between them using the generalized radial basis function (GRBF) kernel as

$$\boldsymbol{y}_m^k = \sum_{l=1}^{N_c} w_l^k \phi\left(||\boldsymbol{x}_m - \boldsymbol{z}_l||\right) + [1 \;\; \boldsymbol{x}_m] b_l, \qquad (4)$$

where $\{\boldsymbol{z}_l | l = 1, ..., N_c\}$ are N_c kernel centers on the view manifold, w_m^k are the target specific weights of each kernel and b_l is the mapping coefficient of the linear polynomial $[1 \; \boldsymbol{x}_m]$. This mapping can be written in a matrix form

$$\boldsymbol{y}_m^k = \mathbf{B}^k \psi(\boldsymbol{x}_m), \qquad (5)$$

where $\mathbf{B}^k$ is a $d \times N_c$ linear mapping corresponding to target k and composed of the weight terms w_l^k in (4) and where $\psi(\boldsymbol{x_m}) = [\phi(||\boldsymbol{x_m} - \boldsymbol{z}_1||, \cdots, \phi||\boldsymbol{x_m} - \boldsymbol{z}_{N_c}||, 1, \boldsymbol{x_m})]$ is a non-linear kernel mapping that contains the regularization term $[1 \; \boldsymbol{x}_m]$. Since $\phi(x_m)$ is dependent only on the view angle we reason that the identity related information is contained within the term $\boldsymbol{B}^k$. For K training targets, we may obtain their corresponding mapping functions $\mathbf{B}^k$ for $k = \{1, 2, ..., K\}$ that can be stacked together to form a tensor $\mathbf{C} = [\mathbf{B}^1 \; \mathbf{B}^2 \; \ldots \; \mathbf{B}^K]$ that contains identity-dependent information pertaining to different poses. Application of the high-order singular value decomposition (HOSVD) to $\mathbf{C}$ abstracts K identity vectors $\boldsymbol{i}^k \in \Re^K$. This decomposition allows us to synthesize the appearance given an identity vector $\boldsymbol{i}^k$ and a view point $\boldsymbol{x}_m$ according to

$$\boldsymbol{y}_m^k = \mathbf{A} \times_3 \boldsymbol{i}^k \times_2 \psi(\boldsymbol{x}_m), \qquad (6)$$

where $\mathbf{A}$ is the core tensor with dimensionality $d \times N_c \times K$ that captures the coupling effect between the identity and view factors and $\times_j$ denotes mode-j tensor product.

The identity vectors from K training targets, i.e., $\{\boldsymbol{i}^k | k = 1, ..., K\}$, may be interpreted as the basis vectors of a latent space for the identity factor. One may think any linear combination of these basis vectors would form a new identity vector, which may lead to a meaningful shape interpolation. However the shape produced in this manner normally does not resemble a real world object. A LD space that is appropriately supported by all training targets is needed to ensure valid shape reconstruction. This motivates us to learn a 1D identity manifold via B-spline curve fitting in the tensor coefficient space $\{\boldsymbol{i}^k | k = 1, ..., K\}$ according to the manifold topology defined in (3). It is expected that an arbitrary identity vector along this identity manifold would be more semantically meaningful due to its proximity to training targets, and that it should support valid shape interpolation. Thus (6) defines a compact generative model for multi-view shape modeling that is controlled by two continuous-valued variables each of which follows its own manifold.

4. Inference Algorithm

We use a graphical model to integrate all the factors into one probabilistic framework as shown in Fig. 3, where three variables are involved, 3D position ($\boldsymbol{X}_t$), camera projection ($\boldsymbol{P}_t$) and identity (I_t). The combination of the first two results in a specific pose on the view manifold (i.e., v_t) along with a scaling factor (s_t) regarding the observed target size. Then the problem of ATR becomes the sequential estimation of the posterior probability $p(I_t, \boldsymbol{X}_t, \boldsymbol{P}_t | \boldsymbol{Z}_t)$. Due to the nonlinear nature of this inference problem, we resort to the particle filtering approach [2] that involves a likelihood function and the dynamics of the three latent variables.

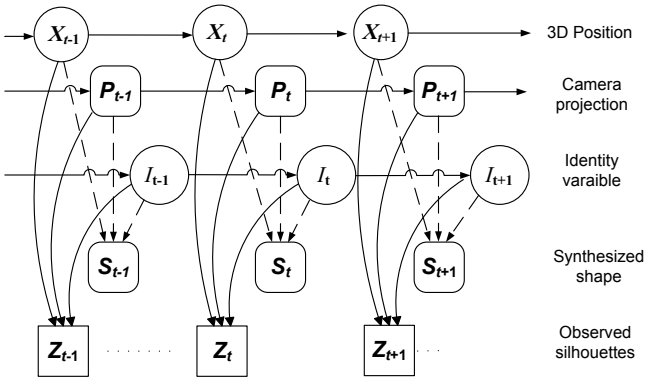


Figure 3. Graphical model for ATR inference.

Given the hypotheses of 3D target position, target identity and camera projection in the tth frame, the hypothesized appearance $\boldsymbol{S}_t$ is reconstructed by the generative model defined in (6) (also scaled by s_t). Then the likelihood function is defined by matching $\boldsymbol{S}_t$ with the observed shape $\boldsymbol{Z}_t$ as

$$p(\boldsymbol{Z}_t | I_t, \boldsymbol{X}_t, \boldsymbol{P}_t) \propto \exp\Big[- \frac{\|\boldsymbol{Z}_t - \boldsymbol{S}_t\|^2}{2\sigma^2} \Big], \qquad (7)$$

where σ^2 controls the likelihood sensitivity and $\| \cdot \|$ gives the mean square error between the observed and hypothesized target shapes. Based on the SENSIAC data set, we chose the maneuvering motion model proposed in [5] for $P(\boldsymbol{X}_t | \boldsymbol{X}_{t-1})$ and assume that the camera is static ($\boldsymbol{P}_t = \boldsymbol{P}$). To ease the inference process, the identity manifold can be mapped to a circle along which we need to specify a dynamic model for identity estimation. It is perceivable that the dynamics of the identity variable should be influenced by the current view/pose due to the fact that in some views (e.g., side views) the identity is much more distinguishable than in other views (e.g., the rear/front/top views). Thus we define a view-sensitive prior $P(I_t | I_{t-1}, v_{t-1})$ that propagates the identity hypothesis adaptively along the identity manifold according the recent view, i.e., v_{t-1}. The most computational demanding step is the likelihood calculation that involves online shape reconstruction from the generative model for each hypothesis during inference.

5. Experimental results

We have developed three particle filtering-based ATR algorithms that share the same inference framework as shown in Fig. 3. Method-I uses the proposed generative target model that involves both the view and identity manifolds for shape interpolation (i.e., both the identity and view variables are continuous.). Method-II uses a simplified version where only the view manifold is involved for shape interpolation (i.e., the identity variable is discrete). Method-III is a traditional template-based method that only uses the training data for shape matching without shape interpolation (i.e., both the view and identity variables are discrete.).

In the following, we first present the learning of the proposed generative model along with some simulations of shape interpolation along the view and identity manifolds. Then we introduce the SENSIAC dataset [1], followed by detailed experimental results regarding the overall ATR performance of the two competing algorithms. Background-subtraction [32] was applied to all test sequences to get the initial segmentation of the target in each frame and the distance transform [4] was also applied to create observation sequences used for ATR inference.

5.1. Generative Model Learning

We acquired 32 3D models of different ground vehicles to learn and evaluate the proposed generative model (8 tanks, 7 APCs, 9 SUVs and 8 pick-ups), as shown in Fig. 4. All 3D models were scaled to have similar size, and targets in the same class shared the same scaling factor. This class-dependent scaling allowed us to estimate the real range information in a 3D scene. For each 3D model, we generated a set of silhouettes corresponding to 150 training view points non-uniformly distributed on the view manifold. These training views are relatively sparser near the top of the view hemisphere and denser near the bottom. This is because there is less distinguishable shape variability in a top-down view compared with that in a profile side view. The distance transform [4] was applied to "soften" these silhouettes. This setup supports learning the generative model jointly with the identity manifold as described in Section 3.

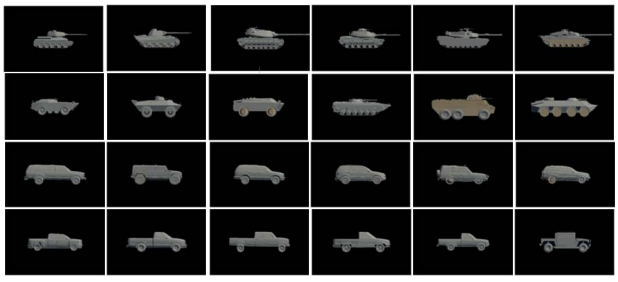
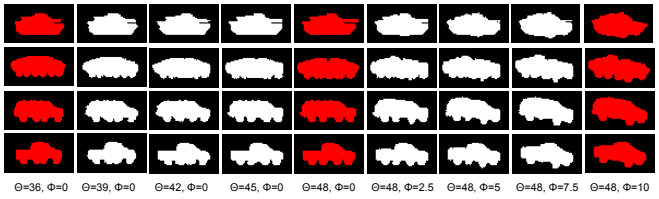
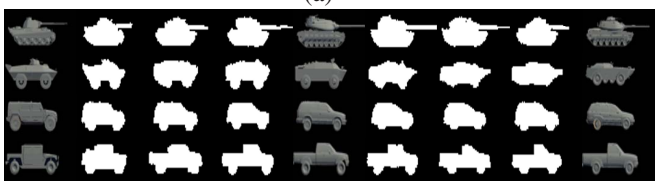


Figure 4. Some of the 3D CAD models used for model learning (from top to bottom: tanks, APCs, SUVs, and pick-ups) ordered according to the manifold topology determined by (3).



(a)



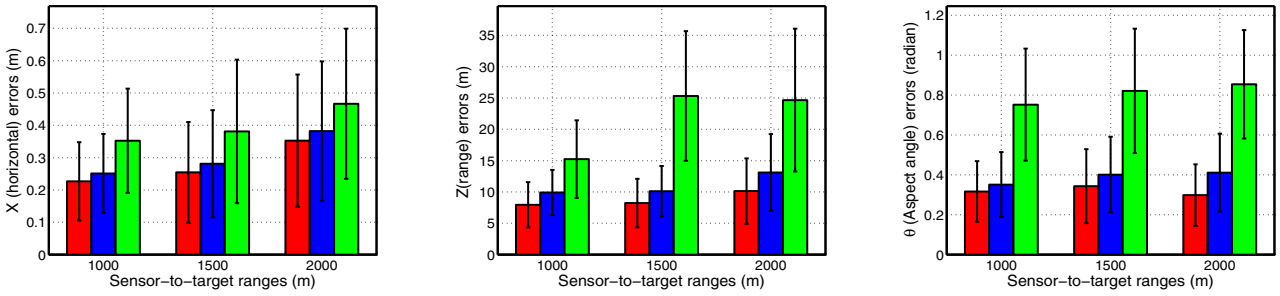
(b)

Figure 5. Shape interpolation along the view and identity manifolds for four target classes. (a) Shape interpolation along the view manifold: the dark shapes are training ones that are adjacent on the view manifold, while the white ones are interpolated. First and second training shapes are 12° apart along the azimuth angle, and the second and third ones are 10° apart in the elevation angle. (b) Shape interpolation along the identity manifold: the dark shapes are training ones that are adjacent on the identity manifold, while the white ones are interpolated.

After model learning, we performed testing to examine the capability of shape interpolation of the learned generative model along the two manifolds. First, we picked a target from each of the four target classes and computed three interpolated shapes (after thresholding) between two training views along both the azimuth angle and the elevation angle, as shown in Fig. 5(a). Second, we tested the identity manifold by computing six interpolated shapes along the identity manifold between three training targets that were adjacent along the identity manifold, as shown in Fig. 5(b). A smooth transition can be observed among the series of interpolated shapes, despite the fact that the training shapes are quite different from one another. Both results show that the generative model is able to interpolate semantically meaningful target shapes along the two manifolds.

5.2. Tests on the SENSIAC database

The SENSIAC ATR database contains a large collection of visible and MWIR (midwave infrared) imagery of seven military and two civilian vehicles. Vehicles drive in a continuous circle marked on the ground with a diameter of 100 meters (m). The imagery was taken at a frame rate of 30Hz for one minute from 1,000m to 5,000m (with a 500m increment) during both day and night time conditions. Specifically, we chose 12 infrared sequences (each has around 1000 frames) of four vehicles (as shown in Fig. 7) acquired at three ranges (1000m, 1500m, and 2000m) during night for algorithm evaluation.



(a) (b) (c)

Figure 6. Overall 3D tracking performance of Method-I (shape interpolation along the view and identity manifolds, the left red error bar), Method-II (shape interpolation along the view manifold, the middle blue error bar), and Method-III (no shape interpolation, the right green error bar) averaged over 12 SENSIAC sequences. (a) Horizontal errors (m). (b) Range errors (m). (c) Aspect angle errors (rad).

Figure 7. Four vehicles used in algorithm evaluation.

Additionally, the SENSIAC database includes a rich set of meta data for each infrared sequence, such as true north offsets of sensors (in azimuth and elevation, Fig.8(a)), the target type, the target speed in each frame, the range and slant ranges from the sensor to the target for each frame (Fig.8(b)), the pixel location of the target centroid in each frame, heading direction with respect to north in each frame, and aspect orientation of the vehicle (Fig.8(c)). Moreover, we defined a sensor-centered 3D world coordinate system (Fig.8(d)) and developed a pinhole-based camera calibration technique that is used to estimate the ground-truth 3D position of the target in each frame according to the meta data. The tracking performance is based on the errors in the estimated 3D position and aspect orientation.

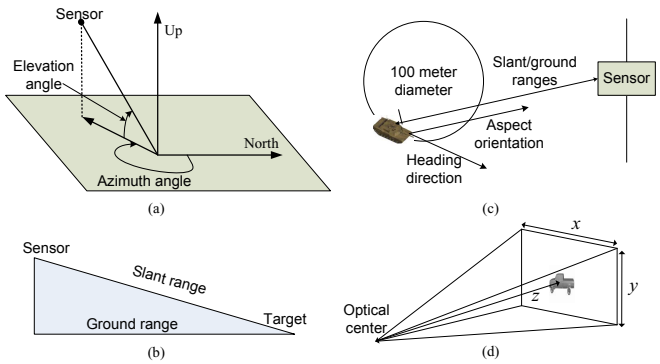


Figure 8. The spatial setting between the sensor and the target. (a) The sensor orientation in the world coordinate system; (b) The slant and ground ranges (side view); (c) The aspect orientation and the heading direction (top-down view); (d) A sensor-centered coordinate system used for algorithm evaluation.

5.2.1 Tracking Evaluation

We computed the errors of estimated 3D positions along x (along the horizon) and z axes (along the range) (Fig.8(d)), as well as that of the aspect orientation of the target (Fig.8(c)). All tracking trials were initialized by the ground-truth information in the first frame and the overall tracking performance is shown in Fig. 6. The three algorithms have achieved comparable results in the horizontal errors (less than one meter), with Method-I delivering performance gains of 10% and $20\% - 40\%$ over Method-II and Method-III, respectively. Method-I also outperformed other two methods on the range and aspect estimation with over $5 - 25\%$ and $50 - 70\%$ improvements, especially when the range was large (i.e., the target was relatively small). We also present some tracking results of Method-I for four 1000m sequences in Fig. 9, where the interpolated shapes are overlaid on the target according to the estimated 3D position and aspect angle as well as the given camera model. All of these results show the usefulness of the generative model in interpolating target shapes for ATR tasks.

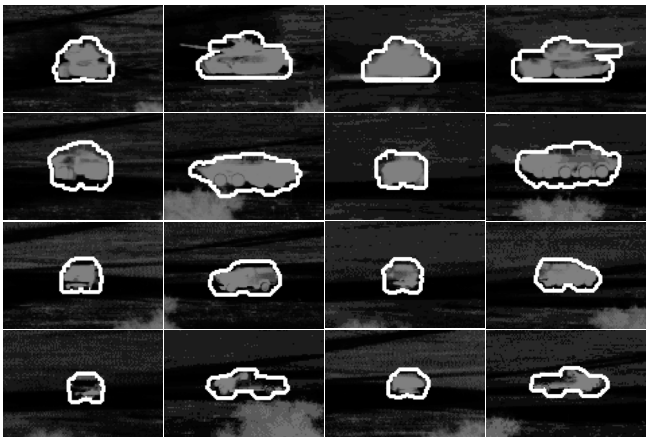


Figure 9. Target tracking results for four 1000m sequences overlaid with interpolated shapes (from top to bottom: T72 tank, BTR APC, ISUZU SUV, Ford pick-up truck).

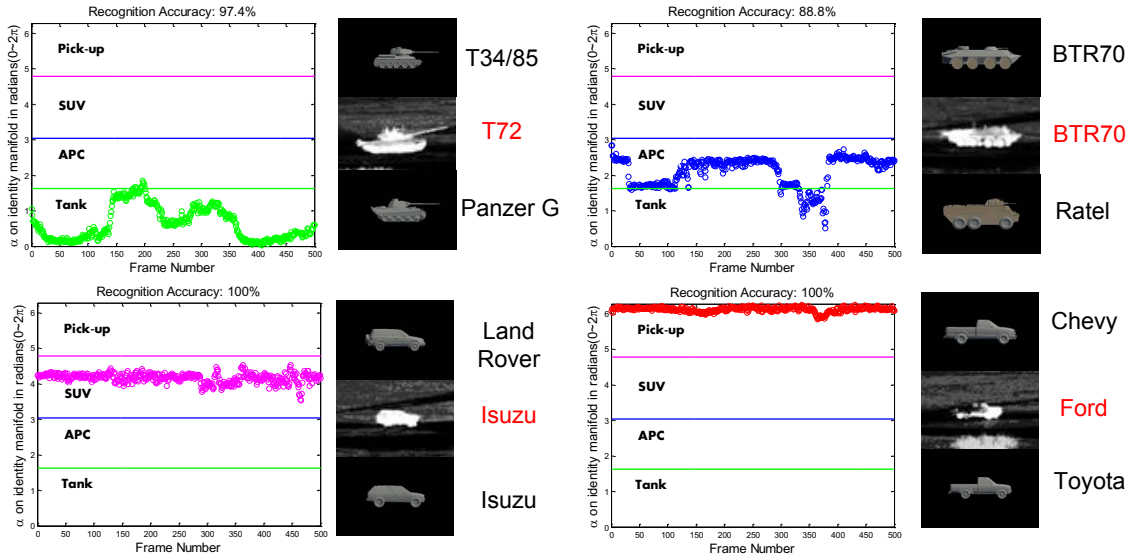


Figure 10. Target recognition results (frame by frame) for four sequences (1000m range) (we down-sampled 1000 frames to 500 frames for display. The vertical axis is the estimated identity in terms of an angular variable ($\alpha \in [0, 2\pi)$) and the ranges of α with respect to four target classes are also shown. Additionally, the actual target type is also given along with the two adjacent training target types.

5.2.2 Recognition Evaluation

As mentioned before, the 1D closed-loop identity manifold learned from the tensor coefficient space can be mapped into a unit circle to ease the inference process, and then the identity variable becomes an angular one $\alpha \in [0, 2\pi)$. Correspondingly, the four target classes (tanks, APCs, SUVs, and pick-ups) can be represented by four angular sections along the circular-shaped identity manifold (as shown in Fig. 1). 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 four classes. Also, it is interesting to check the two best-matched training targets for a given sequence that can be found by finding the two nearest neighbors along the identity manifold when the class of the target was correctly determined. The overall recognition results of three methods for 12 sequences are shown in Table 1. Method-I achieved 100% recognition accuracy for all SUV and Pick-up sequences, mainly because that SUV and pick-up have relatively strong distinguishable shapes compared with other targets. Overall, Method-I shows moderate improvements over other two methods with all accuracies above 80%, showing the usefulness of shape interpolation along the two manifolds in the generative model.

Targets	Tank	APC	SUV	Pick-up
1000m	97/93/92	89/88/84	100/99/100	100/100/100
1500m	92/91/85	85/82/80	100/100/99	100/100/98
2000m	85/85/79	80/79/75	100/100/98	100/98/98

Table 1. Overall recognition accuracies ($\%$) of three methods (Method-I/Method-II/Method-III) for 12 SENSIAC sequences .

We present more detailed recognition results of Method-I for the four 1000m sequences in Fig. 10, which shows not only the frame-by-frame identity estimation results but also the two best-matched training targets. In most frames, the estimated identity values are in the correct range and misclassification usually occurs around the front/rear views when the target is not very distinguishable. Interestingly, the two best matches for the APC and SUV sequences include the correct target model. We do not have the 3D model for the T72 tank and the Ford pick-up in our training set, but their best matches still resemble the actual ones.

5.3. More Discussion

The proposed target model was also tested on some visible-band real-world videos where the targets show wider view changes. Our algorithm was able to accurately track the 3D position and the pose along with the correct identity, given reasonable background subtraction results. Still we consider our work preliminary due to two factors. First, we use the silhouette-based shape representation that requires object segmentation prior to tracking. The background subtraction used here assumes that the camera is not moving. In the case of camera motion, object segmentation becomes necessary that could be challenging. Second, we did not consider the occlusion problem and should be taken into account in target representation. The silhouette-based shape representation is a global feature that is sensitive to the occlusion problem. But this work could be extended to other more salient and robust features (such as SIFT and HOG) thereby making the proposed model promising to more real-world applications.

6. Conclusion

We have presented a new shape-based generative model that incorporates two manifolds for multi-view target modeling. The identity manifold was proposed to capture both inter-class and intra-class shape variability among training targets. The view manifold is designed to be hemisphere-shaped and reflects nearly all possible pose variations. A particle filtering-based ATR algorithm was presented that adopts the proposed target model. The experimental results on the SENSIAC ATR dataset show the advantages of shape interpolation along both the view and identity manifolds.

Acknowledgments

This work was supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under grants W911NF-04-1-0221 and W911NF-08-1-0293 and an OHRS award (HR09-030) from the Oklahoma Center for the Advancement of Science and Technology (OCAST).

References

[1] Sensiac, 2008. https://www.sensiac.org/. 33, 37

[2] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online non-linear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.*, 50(2):174–188, 2002. 36

[3] H. Bülthoff and S. Edelman. Psychophysical support for a 2D view interpolation theory of object recognition. *Proc. of the National Academy of Science*, 89:60–64, 1992. 34

[4] A. Elgammal and C. S. Lee. Seperating style and content on a non-linear manifold. In *CVPR*, 2004. 34, 35, 37

[5] X. Fan, G. Fan, and J. Havilcek. Generative models for maneuvering target tracking. *IEEE Trans. on Aerospace and Electronics Systems*, 46(2):635–655, April 2010. 36

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 33

[7] C. Gosch, K. Fundana, A. Heyden, and C. Schnörr. View point tracking of rigid objects based on shape sub-manifolds. In *ECCV*, 2008. 34

[8] S. Khan, H. Cheng, D. Matthies, and H. Sawhney. 3D model based vehicle classification in aerial imagery. In *CVPR*, 2010. 34

[9] A. Kushal, C. Schmid, and J. Ponce. Flexible object models for category-level 3D object recognition. In *CVPR*, 2007. 33, 34

[10] C.-S. Lee and A. Elgammal. Modeling view and posture manifolds for tracking. In *ICCV*, 2007. 34, 35

[11] M. Leotta and J. Mundy. Predicting high resolution image edges with a generic, adaptive, 3-D vehicle model. In *CVPR*, 2009. 34

[12] Y. Li, L. Gu, and T. Kanade. A robust shape model for multi-view car alignment. In *CVPR*, 2009. 33

[13] J. Liebelt and C. Schmid. Multi-view object class detection with a 3D geometric model. In *CVPR*, 2010. 33, 34

[14] J. Lou, T. Tan, W. Hu, H. Yang, and S. Maybank. 3-D model-based vehicle tracking. *IEEE Transactions on Image Processing*, 14:1561–1569, 2005. 34

[15] D. Lowe. Local feature view clustering for 3D object recognition. In *CVPR*, 2001. 33

[16] X. Mei and H. W. S. K. Zhou. Integrated detection, tracking and recognition for ir video-based vehicle classification. In *ICASSP*, 2006. 33

[17] M. I. Miller, U. Grenander, J. A. Osullivan, and D. L. Snyder. Automatic target recognition organized via jump-diffusion algorithms. *IEEE Trans. Image Process.*, 6(1):157–174, 1997. 33

[18] H. Murase and S. Nayar. Visual learning and recognition of 3D objects from appearance. *Int. J. Comput. Vision*, 14:5–24, 1995. 34

[19] O. Ozcanli, A. Tamrakar, and B. Kimia. Augmenting shape with appearance in vehicle category recogntion. In *CVPR*, 2006. 34

[20] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990. 34

[21] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *Int. J. Comput. Vision*, 66:231–259, 2006. 33

[22] R. Sandhu, S. Dambreville, A. Yezzi, and T. A. Non-rigid 2D-3D pose estimation and 2D image segmentation. In *CVPR*, 2009. 34

[23] R. Savarese S., Fergus and L. Fei-Fei. Multi-view object categorization and pose estimation. In *Computer Vision*, volume 285 of *Studies in Computational Intelligence*. Springer, 2010. 33, 34

[24] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *ICCV*, 2009. 33, 34

[25] J. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12:1247–1283, 2000. 34

[26] A. Toshev, A. Makadia, and D. K. Shape-based object recognition in videos using 3D synthetic object models. In *CVPR*, 2009. 34

[27] Y. Tsin, Y. Gene, and V. Ramesh. Explicit 3D modelling for vehicle monitoring in non-overlapping cameras. In *AVSS*, 2009. 34

[28] S. Ullman. An approach to object recognition: Aligning pictorial descriptions. *Cognition*, 32:193–254, 1989. 34

[29] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Trans Pattern Anal Mach Intell*, 13:992–1006, 1991. 34

[30] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensebles: Tensorfaces. In *ECCV*, 2002. 34

[31] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simeple features. In *CVPR*, 2008. 33

[32] Z. Zivkovic and F. van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27:773–780, May 2006. 37