SIMULTANEOUS TARGET RECOGNITION, SEGMENTATION AND POSE ESTIMATION

Liangjiang Yu [§], *Guoliang Fan* [§ *], *Jiulu Gong* [♯ §] *and Joseph P. Havlicek* [‡]
School of Electrical and Computer Engineering, Oklahoma State University, USA [§]
School of Mechatronics Engineering, Beijing Institute of Technology, China [♯]
School of Electrical and Computer Engineering, University of Oklahoma, USA [‡]

ABSTRACT

We propose a simultaneous target recognition, segmentation and pose estimation algorithm for the infrared ATR task. A probabilistic framework of level set segmentation is extended by incorporating a shape generative model that provides a multi-class and multi-view shape prior. This generative model involves a couplet of a view manifold and an identity manifold for general shape modeling. Then an energy function from the probabilistic level set formulation can be iteratively optimized by a shape-constrained variational method. Due to the fact that both the view and identity variables are explicitly involved in the level set optimization, the proposed method is able to accomplish recognition, segmentation, and pose estimation. Experimental results show that the proposed method outperforms two traditional methods where target recognition and pose estimation are implemented after segmentation.

1. INTRODUCTION

Since the Snakes method was first introduced in [1], it had an enormous impact on the segmentation community. But it had some drawbacks as summarized in [2]. On the other hand, the level set method became more and more popular in the field of image segmentation in recent years. Many earlier work focused on low-level features such as intensity, color, texture, motion, which may not be sufficient to handle images with complicated background/foreground [2]. Therefore, some high-level prior knowledge about the shape of expected objects were introduced. Many efforts have been made to incorporate a shape prior, often represented as a signed distance function, in level set segmentation. The first application of shape priors for level set segmentation was developed in [3], where an additional term is added to the contour evolution equation to drive the embedding function to the most likely shape represented by principle component analysis (PCA). This idea was further strengthened in [4] by directly optimizing the shape-driven level set in a PCA-based linear subspace. Then the use of nonparametric model was developed which assumes that the embedding function is modeled as a Gaussian distribution [5], and also another approach allows non-Gaussian distributions was discussed in [6]. In [7, 8], a nonlinear dimensionality reduction method called the Gaussian Process Latent Variable Model (GPLVM) was used to learn a low dimensional shape space that is applied to constrain the solution space of level set segmentation. However, the GPLVM-based shape space only supports one latent variable explicitly, either identity (different objects under the same pose) or pose (different poses for the same identity).

In this paper, we propose a new shape constrained level set algorithm that integrates recognition, segmentation and pose estimation into one probabilistic framework. This work is inspired by [7, 8] and motivated by the recently proposed shape generative model that involves a couplet of the view-identity manifolds (CVIM) for shape modeling [9]. Specifically, we augment the level set framework proposed in [10] by incorporating the CVIM to provide shape priors. A multi-threaded optimization technique is proposed to produce joint target recognition, segmentation, recognition and pose estimation simultaneously. We compare the proposed method with two traditional implementations where segmentation is performed prior to recognition and pose estimation. Experimental results on a set of infrared imagery show the advantages of the proposed algorithm.

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 W911NF-08-1-0293. * Corresponding author.

2. CVIM-BASED SHAPE MODELING

The CVIM was proposed for infrared ATR in [9], as shown in Fig. 1. It is learned from a set of 2D shapes created by 3D CAD models (6 classes and 6 models for each class) by a nonlinear kernalized tensor decomposition. CVIM involves a hemisphere-shaped view manifold and a closed-loop identity manifold in the tensor coefficient space. Two practical considerations lead to this heuristic simplification of the identity manifold. First, all targets are man-made ground vehicles which have different degrees of similarity, thus a closed structure is more suitable than an open one. Second, a 1D closed loop facilitates statistical inference for identity estimation (i.e., target recognition). Also, a *class-constrained shortest-closed-path method* was proposed to find the optimal topology of the identity manifold which ensures the targets of the same class or of similar shapes will stay closer along the identity manifold (i.e., APCs→SUVs→Min-vans→Sedans→Pick-ups→Tanks→APCs). Due to the continuous nature of two manifolds, CVIM can be used to represent unknown vehicles under arbitrary view point, which is especially desirable for tracking and recognition from image sequences.

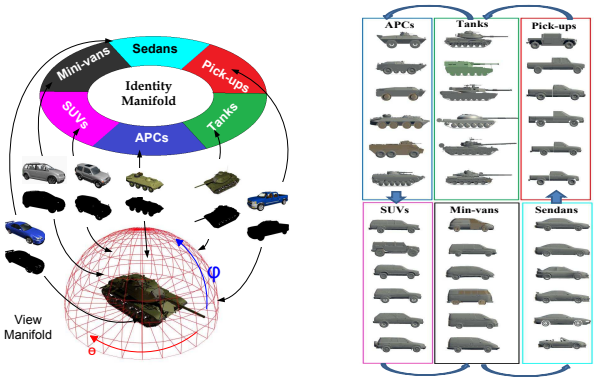

Fig. 1. The CVIM proposed in [9] with 36 CAD models for learning.

3. PROPOSED METHOD

We first present a probabilistic formulation of the proposed shape constrained level set framework. Then we develop a three-stage multi-threaded inference algorithm, where after multi-threaded initialization, the solution is achieved by level set-based shape optimization and CVIM-based shape inference alternately. This section is concluded by a summary of the whole inference algorithm.

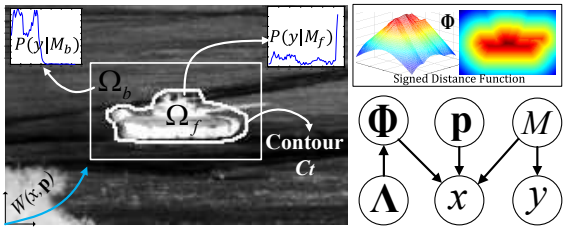

Fig. 2. Left: Representation of a target in an infrared image: the contour C_t, foreground Ω_f and background Ω_b, foreground/background models M, and the warp $W(x,\mathbf{p})$. Top right: the shape embedding function $\mathbf{\Phi}$. Bottom right: the proposed probabilistic framework, where $\mathbf{p}$ is the parameter of a warp function W, x is a pixel location, y is a pixel value. $\mathbf{\Lambda} = \{\alpha, \Theta\}$ where α and Θ are the identity and view variables defined in CVIM respectively .

3.1. Problem Formulation

Implicit contour and level set methods have been proven effective for image segmentation by representing the contour of an object appearing in the scene. The basic idea is to initialize a boundary C and then minimize the energy function related to $\mathbf{\Phi}$ along the negative function gradient. A probabilistic level set segmentation framework was proposed in [10] where an energy function called the pixel-wise posterior was defined to represent an image as a bag of pixels [11] with the background and foreground models obtained from $\mathbf{\Phi}$. In this paper we extend the model from [10] to a new shape-constrained level set segmentation method by introducing $\mathbf{\Lambda} = \{\alpha, \Theta\}$, which represent the view (i.e., pose) and identity variables defined in the CVIM. Adapted from the one in [10], Fig. 2 shows the proposed probabilistic framework which supports joint target recognition, segmentation and pose estimation. Similarly, we can derive a joint probability density function:

$$\begin{aligned} &P(x, y, \mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{p}, M) \\ &= P(x|\mathbf{\Phi}, \mathbf{p}, M)P(y|M)P(\mathbf{\Phi}|\mathbf{\Lambda})P(\mathbf{\Lambda})P(\mathbf{p})P(M) \\ &\propto P(x|\mathbf{\Phi}, \mathbf{p}, M)P(y|M)P(\mathbf{\Phi})P(\mathbf{\Phi}|\mathbf{\Lambda})P(\mathbf{\Lambda})P(\mathbf{p})P(M), \end{aligned} \quad (1)$$

where variables are defined in the caption of Fig.2 and $P(\mathbf{\Phi}|\mathbf{\Lambda})$ involves the template matching to be defined in section 3.4. Similar to [10], by marginalizing over the model M, and using the logarithmic opinion pool, we can derive a new pixel-wise posterior:

$$\begin{aligned} P(\mathbf{\Phi}, \mathbf{\Lambda}, \mathbf{p}|\mathbf{\Omega}) \propto &\prod_{i=1}^{N} \Big\{ \big(P(x_i|\mathbf{\Phi}, \mathbf{p}, y_i)\big) \Big\} P(\mathbf{\Phi}) \\ &\cdot P(\mathbf{\Phi}|\mathbf{\Lambda})P(\mathbf{\Lambda})P(\mathbf{p}), \end{aligned} \quad (2)$$

where $\mathbf{\Lambda}$ is the latent variable of the shape kernel $\mathbf{\Phi}$, and $P(x_i|\mathbf{\Phi}, \mathbf{p}, y_i)$ is defined as:

$$P(x_i|\mathbf{\Phi}, \mathbf{p}, y_i) = H_\epsilon(\mathbf{\Phi}(x_i))P_f + (1 - H_\epsilon(\mathbf{\Phi}(x_i)))P_b, \quad (3)$$

where

$$P_f = \frac{P(y_i|M_f)}{\eta_f P(y_i|M_f) + \eta_b P(y_i|M_b)}, \quad (4)$$

$$P_b = \frac{P(y_i|M_b)}{\eta_f P(y_i|M_f) + \eta_b P(y_i|M_b)}, \quad (5)$$

where η_f and η_b are number of pixels belong to the foreground and background region respectively, $P(y_i|M_f)$ and $P(y_i|M_b)$ are foreground and background models represented by 64 bins histograms, and $H_\epsilon(\cdot)$ is a smoothed Heaviside step function. Here we specify the prior of the shape embedding function $P(\mathbf{\Phi})$ that encourages $\mathbf{\Phi}$ to resemble a signed distance function as [10]:

$$P(\mathbf{\Phi}) = \prod_{i=1}^{N} \Big[\frac{1}{\sigma\sqrt{2\pi}} \exp\Big(- \frac{(|\nabla\mathbf{\Phi}(x_i)| - 1)^2}{2\sigma^2} \Big) \Big]. \quad (6)$$

Then the goal of shape-constrained level set segmentation is to maximize (2) with respect to $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ as:

$$(\mathbf{\Phi}^*, \mathbf{\Lambda}^*) = \arg\max_{\mathbf{\Phi}, \mathbf{\Lambda}} P(\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{p}|\mathbf{\Omega}). \quad (7)$$

Unlike [10], where the level set energy function was optimized (with respect to $\mathbf{\Phi}$ only) by calculus of variations, it may not be straightforward to optimize (2) due to the co-existence of $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ as well as the nonlinear and multi-modal nature of CVIM.

3.2. Multi-threaded Optimization: Initialization

In this work, we propose a multi-threaded optimization algorithm to solve (7), as shown in Fig. 3. The initialization stage has three steps to initialize the multi-threaded optimization that is needed to endure efficient and accurate inference results. First, given a bounding box ($\mathbf{\Phi}_0$), a traditional level set (without shape prior) is used for initial segmentation ($\mathbf{\Phi}_1$). Then, by using the height/width ratio, we can find a small set of the best matched training shapes with known view and identity values in the CVIM. Third, via template matching between the segmented shape and selected training shapes, L most potential candidates ($\mathbf{\Lambda}_0^{(1:L)}$) are selected as the seeds to start the multi-threaded optimization to estimate $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ iteratively.

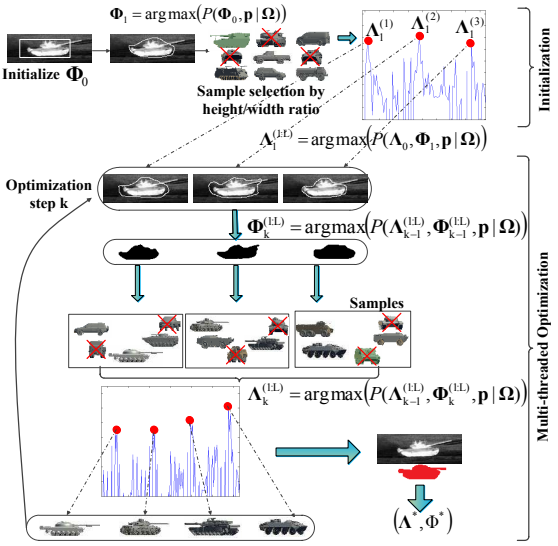

Fig. 3. The illustration of the three-stage inference algorithm.

3.3. Multi-threaded Optimization: Shape Inference

At this stage, we are only looking for a shape contour that maximize the energy function (2) under a shape prior as

$$\mathbf{\Phi}_k^{(l)} = \arg\max_{\mathbf{\Phi}} P(\mathbf{\Lambda}_{k-1}^{(l)}, \mathbf{\Phi}_{k-1}^{(l)}, \mathbf{p}|\mathbf{\Omega}), \tag{8}$$

where k is the iteration index and $l = 1...L$ is the thread index. $\mathbf{\Lambda}_{k-1}^{(l)}$ corresponds a shape prior specified by CVIM that is used to initialize the level set optimization. Substitute (3) and (6) into (8), take the log and then take the first variation with respect to $\mathbf{\Phi}$, the term $P(\mathbf{\Phi}|\mathbf{\Lambda})$, $P(\mathbf{\Lambda})$, $P(\mathbf{p})$ will be dropped, and here we get:

$$\frac{\partial f}{\partial \mathbf{\Phi}} = \frac{\delta_\epsilon(\mathbf{\Phi})(P_f - P_b)}{P(x|\mathbf{\Phi}, \mathbf{p}, y)} - \frac{1}{\sigma^2}[\nabla^2\mathbf{\Phi} - div(\frac{\nabla\mathbf{\Phi}}{|\nabla\mathbf{\Phi}|})], \tag{9}$$

where $\sigma^2 = 50$, $f = \log P(\mathbf{\Phi}, \mathbf{\Lambda}, \mathbf{p}|\mathbf{\Omega})$, ∇^2 is the Laplacian operator and $\delta_\epsilon(\mathbf{\Phi})$ is derivative of a blurred Dirac delta function, and $div(\cdot)$ is the divergence operator [12]. This is similar to the level set shape optimization in [10], and can be optimized by steepest-ascent by gradient flow $\frac{\partial f}{\partial \mathbf{\Phi}} = \frac{\partial f}{\partial t}$, for stability $\frac{\tau}{\sigma^2} < 0.25$ must be satisfied, where τ is the time step [10].

3.4. Multi-threaded Optimization: CVIM Inference

Given a shape embedding function $\mathbf{\Phi}_k$ (where we have dropped the thread index for simplicity), we will optimize $\mathbf{\Lambda}_k$ by performing CVIM inference as

$$\begin{aligned} \mathbf{\Lambda}_k &= \arg\max_{\mathbf{\Lambda}} P(\mathbf{\Lambda}_{k-1}, \mathbf{\Phi}_k, \mathbf{p}|\mathbf{\Omega}), \\ &= \arg\max_{\mathbf{\Lambda}} \{P(\mathbf{\Phi}_k|\mathbf{\Lambda}_{k-1})P(\mathbf{\Lambda}_{k-1})\}, \end{aligned} \tag{10}$$

where

$$P(\mathbf{\Phi}_k|\mathbf{\Lambda}_{k-1}) \propto \exp\big(-\frac{\|\mathcal{C}(\mathbf{\Phi}_k) - \mathcal{S}(\mathbf{\Lambda}_{k-1})\|^2}{2\xi^2}\big), \tag{11}$$

where $\mathcal{C}(\mathbf{\Phi}_k)$ is a shape obtained from $\mathbf{\Phi}_k$, $\mathcal{S}(\mathbf{\Lambda}_{k-1})$ is the CVIM-based shape interpolation given $\mathbf{\Lambda}_{k-1}$, $\|\cdot\|$ represents the shape matching error, and ξ control the sensitivity of shape matching. $P(\mathbf{\Lambda}_{k-1})$ is the prior probability from the previous step. Furthermore, we developed a Markov chain Monte Carlo (MCMC)-based inference algorithm for multi-threaded CVIM inference to optimize (10), which is interleaved with the level set shape optimization defined in (8) iteratively. Fig.4 illustrates the major steps in the MCMC-based CVIM inference.

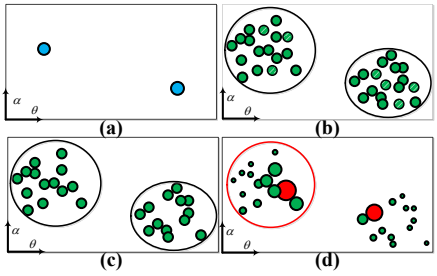

Fig. 4. MCMC-based CVIM inference. (a) The seeds in the latent space for multi-threaded shape estimation. (b) MCMC sampling for each thread. (c) Sample screening based on the height/width ratio. (d) Sample weighting by shape matching and multi-thread reset.

3.5. Algorithm Flow

We summarize the multi-threaded optimization in Algorithm 1 that combines the three stages together.

Algorithm 1 Multi-threaded optimization

Initialization
- Initialize a bounding box $\mathbf{\Phi}_0$ around the object
- Optimize (8) to get $\mathbf{\Phi}_1$ and its height/width ratio (HWR) ($\gamma(\mathbf{\Phi}_1)$)
- Select training shapes with HWR similar to $\gamma(\mathbf{\Phi}_1)$ for template matching
- Initialize CVIM with top L best matched training shapes, $\mathbf{\Lambda}_1^{(1:L)}$ (Fig. 4(a))

for each MCMC iteration ($k = 2...K$) **do**
  for each thread ($l = 1 : L$) **do**
    Shape Inference
    - Initialize a shape prior from previous CVIM inference, $\mathcal{S}(\mathbf{\Lambda}_{k-1}^{(l)})$
    - Optimize (8) to get $\mathbf{\Phi}_k^{(l)}$

    CVIM Inference
    - Optimize (10) by MCMC
    - Draw samples around $\mathbf{\Lambda}_{k-1}^{(l)}$ in the shape space (Fig. 4(b))
    - Discard samples according to $\gamma(\mathbf{\Phi}_k^{(l)})$ (Fig. 4(c))
    - Evaluate the left samples by template matching (11) (Fig. 4(d))
    - Find the local maximum to be new $\mathbf{\Lambda}_k^{(l)}$

  end for
end for

Obtain the final reconition/pose estimation result, $\mathbf{\Lambda}^*$, which is selected from $\mathbf{\Lambda}_K^{(l:L)}$ by finding which one yields the largest level set energy function defined in (2) and $\mathbf{\Phi}^* = \mathcal{S}(\mathbf{\Lambda}^*)$ is the final segmentation result.

4. EXPERIMENTAL RESULT

4.1. Experimental Setup

The CVIM learning is the same as that in [9]. We selected six 3D CAD models for each of the six classes (totally 36 models, as shown in Fig. 1): APCs (Armored Personnel Carriers), Tanks, Pick-ups, Vans, Sedans and SUVs. We considered the elevation angles in the range of $0^\circ \sim 45^\circ$ and azimuth angles $0^\circ \sim 360^\circ$, with 10° and 12° intervals respectively, resulting in 150 multi-view shapes for each target. To reduce the learning complexity, a simple yet efficient DCT-based shape descriptor proposed in [13] was used, where only 10% DCT coefficients are used for CVIM learning. Moreover, this DCT-based shape representation can be used to reconstruct a shape at arbitrary magnification factors by appropriately zero-padding DCT coefficients prior to inverse DCT, avoiding additional zooming or shrinking operations to accommodate the scaling factor.

In addition to the proposed one (referred to as Method-III), we have developed two traditional implementations where target segmentation is performed prior to pose estimation and target recognition. The first one (Method-I) applies background subtraction [14] that is only suitable for the case of a stationary camera, and the second one (Method-II) uses level set segmentation without shape prior [10]. Method-I and Method-II only involve the multi-threaded optimization for CVIM inference (Section 3.4) without shape optimization. The three algorithms were evaluated against the 24 midwave IR sequences from the SENSIAC ATR Database [15], which include 23 night-time and one day-time IR imagery of eight civilian and military ground vehicles moving around a closed circular path with a diameter of 100 meters at three ranges from 1km, 2km and 3km. We selected 100 frames by down-sampling each sequence that is 1800 frames originally, where the aspect angle ranges from 0° to 360° with a $5^\circ - 10^\circ$ interval.

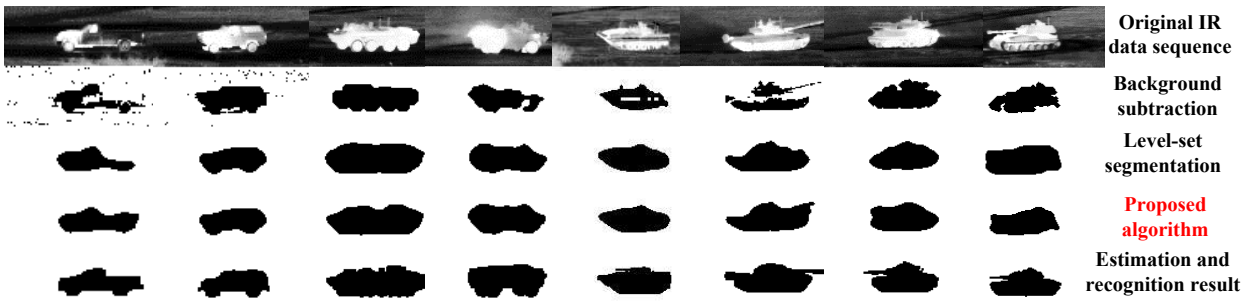

Fig. 5. Snapshot of the segmentation results. From the first row to the last: Original IR frame, background subtraction results, level set segmentation without prior shape, the proposed method, and the final pose estimation and recognition result interpolated from CVIM.

Table 1. Overall segmentation, pose estimation and recognition results of three methods (Method-I/Method-II/Method-III).

Results /ranges	Aspect angle error $\theta(^\circ)$	2D Pixel location error (pixels)	Overlap (%)	recognition accuracy (%)	Range error (meters)
1000m	17.2 / 17.9 / **15.1**	2.8 / 3.1 / **1.9**	85.2 / 82.9 / **88.1**	81 / 78 / **85**	25.1 / 27.8 / **24.2**
2000m	21.2 / 25.2 / **18.7**	2.9 / 3.4 / **2.3**	75.6 / 74.1 / **79.5**	71 / 64 / **73**	39.1 / 38.2 / **33.8**
3000m	26.1 / 27.5 / **21.7**	2.5 / 3.8 / **2.2**	67.7 / 65.5 / **70.1**	69 / 62 / **70**	43.5 / 48.3 / **40.2**

SENSIAC data also provide a rich set of meta data for performance evaluation, such as the aspect angle of the target, the field of view, the $2D$ bounding box of the target in each frame. We will evaluate all three algorithms with respect to the accuracy of pose estimation (i.e., the aspect angle), the 2D pixel location error between the segmented shape and the ground truth bounding box and the overlap ratio between the area they covered, the recognition accuracy in terms of six major target classes, and the sensor-target distance in meters (assuming the real 3D dimension is known for each target type). We will also evaluate the capability of the proposed algorithm for sub-class recognition, i.e., the specific target type within a class.

4.2. Performance Evaluation

Fig.5 shows some snapshots of original IR imagery of eight targets under the 1km range, along with the segmentation results of background subtraction, level set segmentation without shape prior, and the proposed algorithm, followed by the final recognition/pose estimation result in the last row. We can see how the CVIM-based shape prior drives the level set segmentation result to a more semantically meaningful shape, which further enhances the accuracy of pose estimation and identity recognition. In Table 1, some numerical results of the three different methods are shown. We can see clearly that Method-III can achieve moderate and significant improvements over Method-I and II, respectively. Although the improvement is not significant compared with Method-I, the prerequisite of background subtraction makes it less practical in reality. Large errors usually occur at the frames where the target is observed with significant ambiguity and uncertainty from a near frontal or rear view or under strong clutter (such as smoke and dust). Our current implementation is based on an un-optimized MATLAB code, and it is about 20-30 seconds per frame on a dual Core PC desktop computer (2.6GHz CPU and 5G memory). The main computational cost (80%) is from MCMC-based CVIM inference. We are currently developing a gradient-based optimization method that is expected to significantly enhance the algorithm efficiency.

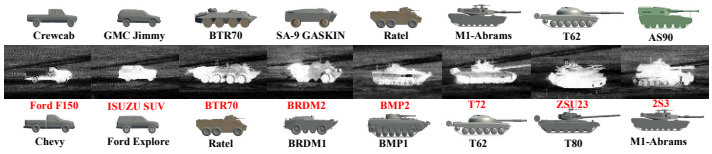

Fig. 6. Sub-class recognition result under 1km. The first row and third row are the closest training vehicle along the identity manifold, the middle row is the original IR data.

Fig.6 shows some sub-class recognition results for eight 1km infrared images. The sub-class recognition can be achieved by finding the two closest training target types along the closed-loop identity manifold in CVIM. Since only the BTR70 model is included in the training data, we find that we can recognize BTR70 at the sub-class level. Interestingly, we can see that T72, BMP2, and 2S3 are also recognized as similar vehicles in our training data: T80, BMP1, and AS90 respectively, showing the usefulness of shape interpolation along the identity manifold.

5. CONCLUSION

In this paper we have integrated a shape-based generative model, i.e., CVIM, into a level set probabilistic level set framework to implement joint target recognition, segmentation and pose estimation. We also implemented a MCMC-based multi-threaded optimization algorithm to infer the shape and two shape-related latent variables (identity and view) in CVIM. Experimental results on a set of infrared imagery demonstrate the advantages of the proposed algorithm over other two traditional implementations where target segmentation is performed prior to recognition and pose estimation.

6. REFERENCES

[1] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Intl. J. of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[2] D. Cremers, M. Rousson, and R. Deriche, "A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape," *Intl. J. of Computer Vision*, vol. 72, no. 2, pp. 195–215, 2007.

[3] M.E. Leventon, W.E.L. Grimson, and O. Faugeras, "Statistical shape influence in geodesic active contours," in *Proc. CVPR*, 2000.

[4] A. Tsai, Jr. Yezzi, A., W. Wells, C. Tempany, D. Tucker, A. Fan, W.E. Grimson, and A. Willsky, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE Trans. Medical Imaging*, vol. 22, no. 2, pp. 137 –154, feb. 2003.

[5] M. Rousson and N. Paragios, "Shape priors for level set representations," in *Proc. ECCV*. 2002, pp. 78–92, Springer.

[6] D. Cremers, S. Osher, and S. Soatto, "Kernel density estimation and intrinsic alignment for shape priors in level set segmentation," *Intl. J. of Computer Vision*, vol. 69, no. 3, pp. 335–351, Sept. 2006.

[7] V. Prisacariu and I. Reid, "Nonlinear shape manifolds as shape priors in level set segmentation and tracking," in *Proc. CVPR*, 2011.

[8] V. Prisacariu and I. Reid, "Shared shape spaces," in *Proc. ICCV*, 2011.

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

[10] C. Bibby and I. Reid, "Robust real-time visual tracking using pixel-wise posteriors," in *Proc. ECCV*, 2008.

[11] T. Jebara, "Images as bags of pixels," in *Proc. ICCV*, 2003.

[12] C. Li, C. Xu, C. Gui, and M.D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE Trans. Image Processing*, vol. 19, no. 12, pp. 3243–3254, 2010.

[13] V.A. Prisacariu and I. Reid, "Shared shape spaces," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, nov. 2011, pp. 2587 –2594.

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

[15] "Military sensing information analysis center (sensiac)," 2008, https://www.sensiac.org/external/index.jsf.