# HYBRID MACHINE LEARNING APPROACHES FOR SCENE UNDERSTANDING: FROM SEGMENTATION AND RECOGNITION TO IMAGE PARSING

By

LIANGJIANG YU

Bachelor of Science in Communication Engineering
North University of China
Taiyuan, Shanxi, China
2008

Master of Science in Electrical Engineering
Oklahoma State University
Stillwater, Oklahoma
2011

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2017

# HYBRID MACHINE LEARNING APPROACHES FOR SCENE UNDERSTANDING: FROM SEGMENTATION AND RECOGNITION TO IMAGE PARSING

Dissertation Approved:

Dr. Guoliang Fan

Dissertation Advisor

Dr. Martin Hagan

Dr. Weihua Sheng

Dr. Christopher John Crick

# ACKNOWLEDGMENTS

I would like to express my sincere gratitude and deepest appreciation to all those who have helped me to make this dissertation possible.

First, I would like to thank my advisor, Dr. Guoliang Fan for his invaluable guidance and support throughout my graduate studies. His inspiring ideas pave the path to make this dissertation happen. His insightful suggestions, inspirational encouragement helped me to overcome many difficulties. His enthusiasm and dedication to research benefit not only my current studies, but also my future careers.

I am very grateful to my committee members, Dr. Martin Hagan, Dr. Weihua Sheng, and Dr. Christopher John Crick for their dedication and commitment throughout the review of this dissertation, I greatly appreciate their extremely helpful guidance during the past few years. I would also like to thank Dr. Damon Chandler for his incredibly helpful comments and support.

I owe my great gratitude to Dr. Joseph Havlicek for his time and efforts on reviewing my paper. I would like to thank Dr. Vijay Venkataraman for his invaluable suggestions and brilliant ideas. Many thanks to Dr. Jiulu Gong for his insightful ideas, and thank you to Dr. Yi Ding, Dr. Xin Zhang for their kind support. I am thankful to Dr. Song Ge and Dr. Meng Ding for working together with me in the lab for several years. Thank you to all VCIPL lab members including Jie Yuan, Jung Jae Yin, Dr. M.N.A. Khan, Ayesha Siddiqua, Mahdi Yazdanpour, Lin Guo, etc.

Last but not the least, I appreciate all my family members for their continuous and unconditional support. I can hardly complete this work without their understanding and encouragement.

This work was supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under Grant W911NF-08-1-0293, the Oklahoma Center for the Advancement of Science and Technology (OCAST) under Grant HR09-030 and Grant HR12-30 and the National Science Foundation (NSF) under Grant NRI-1427345.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: LIANGJIANG YU

Date of Degree: MAY, 2017

Title of Study: HYBRID MACHINE LEARNING APPROACHES FOR SCENE UNDERSTANDING: FROM SEGMENTATION AND RECOGNITION TO IMAGE PARSING

Major Field: ELECTRICAL ENGINEERING

Abstract: we alleviate the problem of semantic scene understanding by studies on object segmentation/recognition and scene labeling methods respectively. We propose new techniques for joint recognition, segmentation and pose estimation of infrared (IR) targets. The problem is formulated in a probabilistic level set framework where a shape constrained generative model is used to provide a multi-class and multi-view shape prior and where the shape model involves a couplet of view and identity manifolds (CVIM). A level set energy function is then iteratively optimized under the shape constraints provided by the CVIM. Since both the view and identity variables are expressed explicitly in the objective function, this approach naturally accomplishes recognition, segmentation and pose estimation as joint products of the optimization process. For realistic target chips, we solve the resulting multi-modal optimization problem by adopting a particle swarm optimization (PSO) algorithm and then improve the computational efficiency by implementing a gradient-boosted PSO (GB-PSO). Evaluation was performed using the Military Sensing Information Analysis Center (SENSIAC) ATR database, and experimental results show that both of the PSO algorithms reduce the cost of shape matching during CVIM-based shape inference. Particularly, GB-PSO outperforms other recent ATR algorithms, which require intensive shape matching, either explicitly (with pre-segmentation) or implicitly (without pre-segmentation). On the other hand, under situations when target boundaries are not obviously observed and object shapes are not preferably detected, we explored some sparse representation classification (SRC) methods on ATR applications, and developed a fusion technique that combines the traditional SRC and a group constrained SRC algorithm regulated by a sparsity concentration index for improved classification accuracy on the Comanche dataset. Moreover, we present a compact rare class-oriented scene labeling framework (RCSL) with a global scene assisted rare class retrieval process, where the retrieved subset was expanded by choosing scene regulated rare class patches. A complementary rare class balanced CNN is learned to alleviate imbalanced data distribution problem at lower cost. A superpixels-based re-segmentation was implemented to produce more perceptually meaningful object boundaries. Quantitative results demonstrate the promising performances of proposed framework on both pixel and class accuracy for scene labeling on the SIFTflow dataset, especially for rare class objects.

## TABLE OF CONTENTS

Chapter Page

# LIST OF TABLES

## LIST OF FIGURES

Figure Page

# CHAPTER I

# INTRODUCTION

Typical configurations of objects appearing in a scene has been studied over the past years in both psychology and computer vision [16]. Many experiments has suggested that precise and/or specifically detailed analysis of individual objects might be enhanced with scene information [17–19], and [20] established cognitive experiments to demonstrate that scene information and meaningful context play an important role in perceptual recognition because objects are almost always perceived in some setting or context. Object recognition tasks are typically conducted by matching appearances of interested objects with available feature libraries while eliminating the background noises. However, background information and surrounding objects in a scene might provide a richer source of knowledge and collection of contextual associations to assist the detection and recognition [21]. In this dissertation, we will mitigate the problem of fully understanding a given scene by studies on both typical object segmentation/recognition algorithms, and scene labeling methods respectively. In this chapter we will briefly introduce the background of automatic target recognition and image parsing techniques.

## *i. Automatic Target Recognition*

Automatic target recognition (ATR) systems perform detection and recognition of extended targets by processing a sequence of images acquired from a passive imaging infrared (IR) sensor [22, 23]. Our interest is primarily in sensors operating in the traditional 3–5 μm mid-wave IR (MWIR) or 8–12 μm long-wave IR (LWIR) bands, although our results are also applicable to those operating in the near, short-wave or far-IR bands, as well. The main functions typically performed by practical IR ATR systems (Figure 1.1) of these types include detection, segmentation, feature extraction, tracking and recognition [2, 24]. While these functions have historically been implemented sequentially, there is a growing recent interest in performing them jointly, so that tracking and recognition are not delayed by ambiguities in the detection process and so that inferences made by the track processor can be leveraged for both recognition and detection.

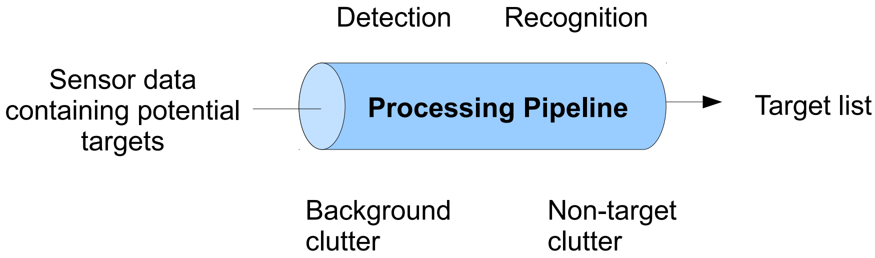


Figure 1.1: Conceptual data flow in automatic target recognition (ATR) systems [2]. Detection algorithms are applied to locate potential targets, and recognition algorithms are implemented at the region of interest for classification.

The infrared ATR problem presents significant challenges. Growth and processing techniques for IR detector materials, such as HgCdTe and InSb, are less mature than those for silicon, and hence, imaging IR sensors are typically characterized by higher noise and poor uniformity compared to their visible wavelength counterparts. The imagery acquired under practical field conditions often exhibits strong, structured clutter, poor target-to-clutter ratios and poor SNR. In important surveillance, security and military applications, the targets of interest may be non-cooperative, employing camouflage, decoys, countermeasures and complex maneuvers in an effort to evade detection and tracking. These difficulties are often exacerbated by the strong ego-motion of the sensor platform relative to the target. Depending on the operational waveband of the sensor, environmental conditions, such as smoke, haze, fog and rain, can result in degraded target signatures, as well as partial or full occlusions. All of these factors contribute to substantial appearance variability of the target thermal signature observed by the sensor, thereby limiting the effectiveness of approaches based on, e.g., stored libraries of static *a priori* signatures. A few examples of MWIR signature variability from the Military Sensing Information Analysis Center (SENSIAC) ATR Algorithm Development Image Database [3] are shown in Figure 1.2. Moreover, one would ideally like the ATR system to be capable of generalizing on the fly, so that both unknown target types and previously unseen views of known target types can be detected, tracked and recognized, at least to within an appropriate target class.

A very large number of ATR algorithms have been proposed in recent decades [2, 24–26]. Some have been based primarily on the computation of certain types of features, such as PCA [27], edge and corner descriptors [28], wavelets [29] or deformable templates [30], while others have been driven more by a particular classification scheme, e.g., neural networks [31], support vector machines (SVM) [32] or sparse representations [4].

On the other hand, in the closely-related fields of computer vision and visual tracking, there have been significant developments in object detection and recognition based on visual features, including the histogram of oriented gradients (HOG) [33,34], the scale-invariant feature transform (SIFT) [35], spin images [36], patch features [37], shape contexts [38], optical flow [39] and local binary patterns [40]. Several feature point descriptors for long-wave IR data applications were evaluated in [41], including SIFT, speeded up robust features (SURF), binary robust invariant scalable keypoints (BRISK), binary robust independent elementary features (BRIEF), fast retina keypoint (FREAK), oriented features from accelerated segment test (FAST) and rotated BRIEF (ORB) features. Certain geometric, topological and spectral descriptors have been widely used, as well [42, 43]. Active contour methods [44, 45] and level set algorithms have also been widely used in shape-based segmentation algorithms [46–48].

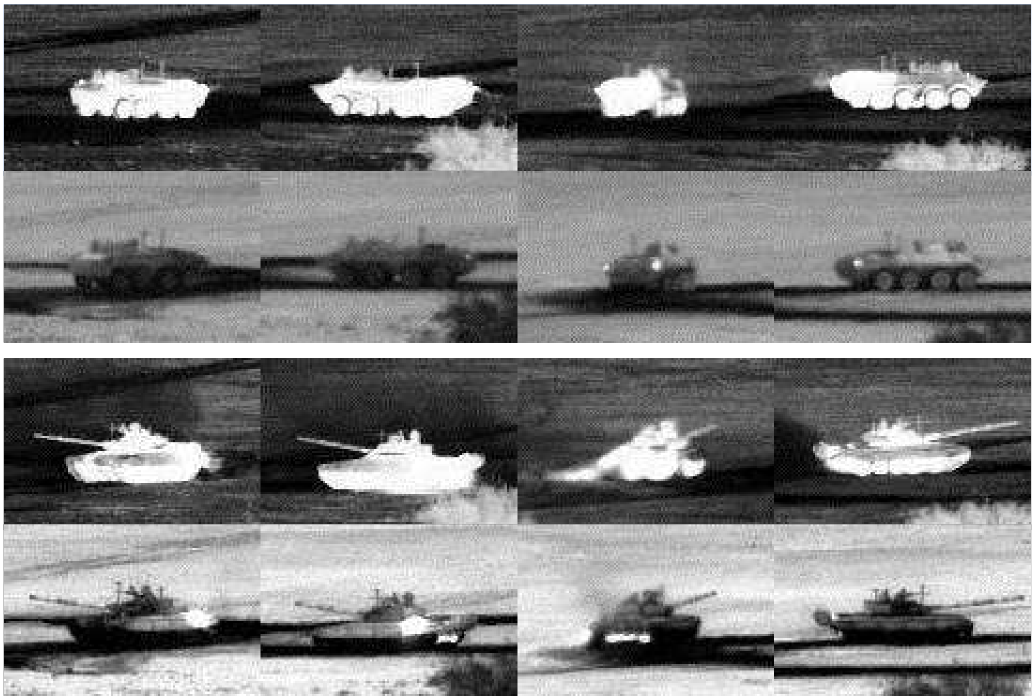


Figure 1.2: Examples of target signature variability from the Military Sensing Information Analysis Center (SENSIAC) ATR database [3]. The first and second rows show diurnal and nocturnal mid-wave IR (MWIR) images of a BTR70 personnel carrier, respectively. The third and fourth rows are diurnal and nocturnal images of a T72 main battle tank. Targets in each column are under the same view.

Shape priors were incorporated into both active contours and level set methods to handle cases of complicated background/foreground structure or objects in [49,50]. In [9], a couplet of identity and view manifolds (CVIM) was proposed for shape modeling by generalizing nonlinear tensor decomposition in [51]. CVIM explicitly defines view and identity variables in a compact latent space and was used with particle filtering for IR tracking and recognition in [9]. Gaussian process latent variable models (GPLVMs) were also used to learn a shape prior in order to accomplish joint tracking and segmentation in [7,52], and GPLVM was further extended for IR ATR application in [53].

In this dissertation, we propose a new shape-constrained level set algorithm that incorporates the parametric CVIM model in a probabilistic framework for integrated target recognition, segmentation and pose estimation. Specifically, we augment the level set framework proposed in [10] by incorporating the couplet of view and identity manifolds (CVIM) [9] which not only provides plausible shape priors, but also eases the process of ATR inference due to the explicit definition of view and identity variables in a compact latent space. The objective energy function of the level set is defined by associating CVIM with observations via a graphical model. To cope with the multi-modal property of CVIM for implicit shape matching, we first develop a particle swarm optimization (PSO) strategy [54] to optimize the energy function with respect to CVIM parameters, and then, we further propose a gradient-boosted PSO (GB-PSO) inspired from [55] to improve the computational efficiency by taking advantage of the analytical nature of the objective function. There are two main contributions. The first one is a unified probabilistic level set framework that integrates CVIM-based implicit shape modeling and naturally supports multiple ATR tasks in one computational flow. The second one is an efficient GB-PSO algorithm that combines both gradient-based and sampling-based optimization schemes for CVIM-based implicit shape matching. Experimental results on the SENSIAC ATR database demonstrate the performance and computational advantages of the proposed GB-PSO [56] over other CVIM-based implementations [57], as well as recent ATR algorithms.

*ii. Image Parsing*

Image parsing, or scene labeling is one of the most important problems of computer vision that pave the path to a more comprehensive scene understanding, and several systems have been developed during the past few years [58–63]. The goal of image parsing is to fully assign semantic labels such as mountain, sea, sky, tree, person, etc.

to each pixel in the observed images, resulting in the recognition and segmentation of objects in both foreground and background in an image, so that every region in the given image is tagged and profiled (Figure 1.3). This detailed and specific description of images gives not only the global understanding of the whole image, but also the semantic location and relation of objects appearing in the scene. Generally we can frame this problem as learning a mapping from 2D pixel grids to the semantic labels, which consists of two parts, feature extraction and inference [64]. Usually we consider descriptive information such as color, texture etc. that are extracted from local patch under various conditions, and we aim to predict the labels of pixels using extracted features during the inference by integrating contextual or spacial dependencies. Since it is relatively hard to distinguish pixels using only low level features (intensity, texture, etc.), so in general cases human tend to recognize regions by spatial correlations and surrounding areas of objects [65]. For example, it is not easy to tell the difference between mountain and forest purely by looking at the local image patches without the global or spatial dependencies. And more interestingly, some conclusive demonstration has shown that performances of human in labeling a small region of an image is not as good as computers [66,67]. To address this challenge, both Markov random fields (MRF) [68, 69] and conditional random fields (CRF) [70–74]

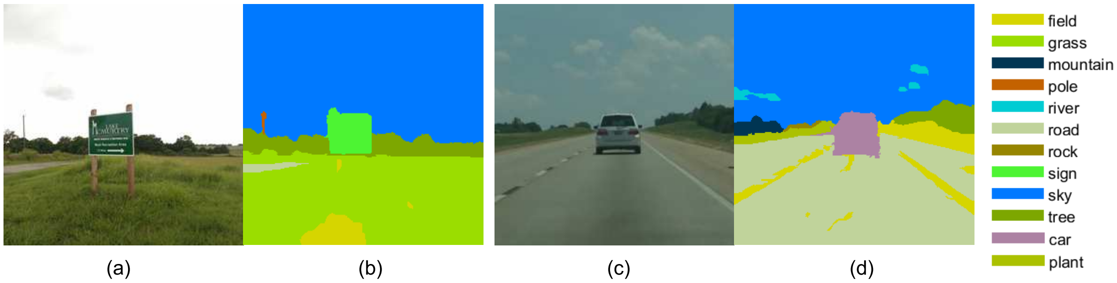


Figure 1.3: Image parsing: fully assign semantic labels to each pixel in the observed image. (a) and (c), input images; (b) and (d), semantic segmentation and labeling maps.

have been used to find high level representation based on spatial dependencies using graphic models, which construct coherence of objects from neighboring regions. However these graphical-model based methods usually suffer from higher computational cost at the prediction step [75]. In recent years the outstanding performances of deep convolutional neural network (CNN) [76] has been incorporated into learning effective and discriminative high-level visual representations [12, 77, 78], which successfully advanced CNNs into various applications, for example, object recognition [79], robotics [80], segmentation [81], detection [82, 83], scene labeling [13], and so on. In [84] a convolution network based scene parsing system was proposed to operate on large input window to produce pixel-wise label hypotheses, where raw image pixels were passed into the convolution network and the supervised training process were implemented from fully-labeled images. And a multi-scale structure was incorporated to ensure contextual consistency. [75] designed a recurrent architecture of CNN from a sequential sets of networks sharing the same set of parameters, where each instance of network was fed with the input image and predictions of the previous network, during which there is no engineered features required.

Non-parametric methods are an alternative to learning-based methods, which involve image retrieval steps in order to semantically find relevant subset of images from the training dataset by global contextual information, and then a graphical model (e.g. MRF, CRF) could be applied to ensure global and neighboring consistency of label transfers [85] from retrieved training images that are similar to the test ones [86–92]. There are two motivations behind these non-parametric approaches [93], the first one is that compared to a large number of labels contained in the training set, usually each image has only limited number of objects appearing in the scene, while the second one is that the large invariance of objects are hard to fit into unified models.

Moreover, many hybrid and/or fusion techniques have been demonstrated to achieve state-of-the-art performances. [94] developed a framework with cascaded classifiers coupled by input and output variables to improve performances. A combination model of motion and appearance features was proposed in [95] for road scenes segmentation and labeling. In [96] a fusion framework was proposed using likelihood from several probabilistic classifiers to achieve better overall accuracy. [13] developed a hybrid framework which alleviated global ambiguity by incorporating a nonparametric label transfer algorithm into a parametric deep CNN model, while learned CNN features were adopted instead of human engineered features for image retrieval and label transfer.

In this dissertation we aim to extend the framework presented in [13] to be more focused on rare classes (e.g. person, boat, car, etc.) that have fewer appearance frequency in the whole training dataset [15]. Specifically, an extension of rare class patches were added to enlarge the retrieval sets for a more balanced exemplar sets during the label transfer [93], but rather than random sampling rare classes over all training images, we added a scene level regulation to constrain the sampling rate for rare classes based on global image scene information, where not only feature similarities and 2D locations were considered, but also sizes of objects were evaluated to produce a more semantically meaningful quantization on the global class likelihood for each pixel. Furthermore, we added another CNN structure which has more emphasis on image patches from rare classes in the training data to enhance the final labeling results. In order to produce a more perceptually meaningful object boundaries, we adopted an efficient graph-based image segmentation [14] to refine the segmentation by counting the majority of labels appearing in each superpixel.

The remainder of this dissertation is organized as follows. Related work including model-based and data-based methods for object recognition, and scene labeling are reviewed in Chapter II. In Chapter III we will first introduce some preliminary work on an probabilistic level-set framework, and the coupled view and identity manifolds (CVIM) with an advanced discrete cosine transform (DCT) formulation. Then we will present the proposed segmentation and recognition framework including analytical formulation, MCMC-based interleaved optimization method, and two PSO-based methods. Chapter IV explores some data-driven methods including the original sparse representation classification (SRC), the multi-attribute Lass with group constraint (MALGC), and also the proposed switchable sparse representation classification (SSRC) method for an improved classification results using sparsity concentration index. In Chapter V we will give a brief introduction on deep learning, convolutional neural network, and a scene labeling framework using integrated parametric and nonparametric model, and in the later part of this chapter we will present the proposed rare class-oriented scene labeling framework (RCSL) with scene information assisted rare classes retrieval method, and a further enhanced superpixels-based re-segmentation method (RCSL-Seg) will be discussed as well. We will summarize the conclusion and future work in Chapter VI.

# CHAPTER II

# RELATED WORK

In this chapter we will briefly introduce past and recent research on object recognition, segmentation and scene understanding, followed by the motivation and contribution of our present work relative to the existing methods.

## 2.1 Automatic Target Recognition and Segmentation

We first categorize several recent ATR algorithms into two main groups, as shown in Figure 2.1.

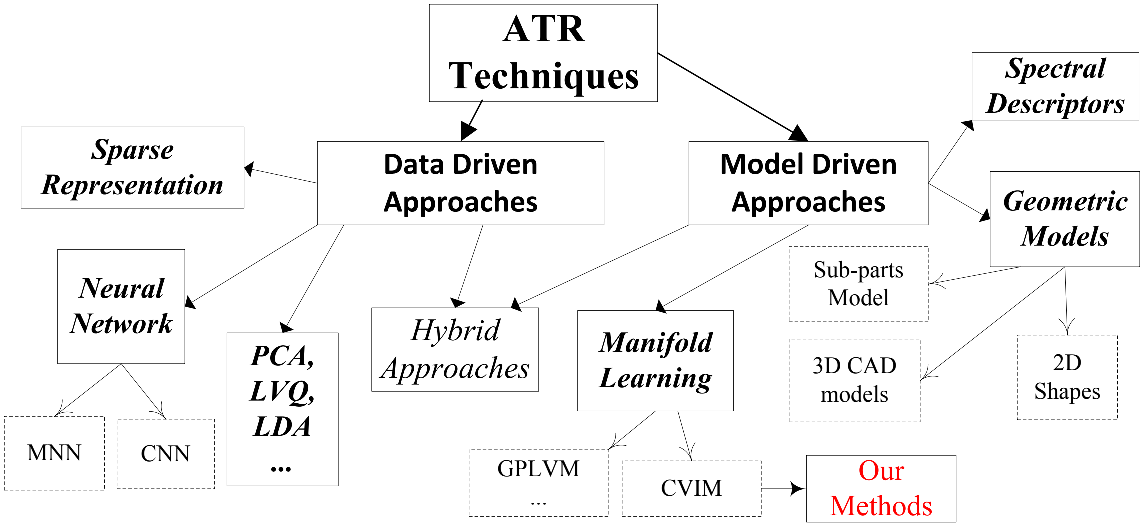


Figure 2.1: A taxonomy of ATR methods. MNN (modular neural network); CNN (convolutional neural network); LVQ (learning vector quantization); GPLVM (Gaussian process latent variable model); CVIM (couplet of view and identity manifolds).

### *i. Data-Driven Approaches*

Data-driven approaches are typically based on learning from a set of labeled real-world training data. Neural networks (NN) [97–101] are an important exemplar. With the NN-based approach, the images acquired from the sensor are treated as points in a high-dimensional vector space. The objective is then to train a large multi-layer perceptron to perform the required mapping from this space to the space of labeled training images [97]. Convolutional neural networks (CNNs) generalize the basic idea by incorporating local receptive fields, weight sharing and spatial sub-sampling to accommodate some degree of translation and local deformation [98]. Modular neural networks (MNNs) are another important generalization where a collection of several independently trained networks each make a classification decision based on local features extracted from a specific region of the image [99]. These individual decisions are then combined to arrive at the overall final ATR classification decision. Another data-driven approach is the vector quantization (VQ)-based method developed in [100, 101], where each target class is trained by the learning vector quantization

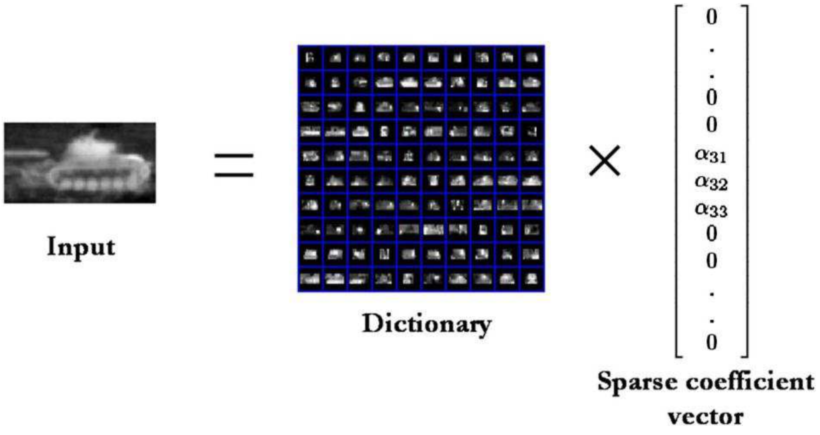


Figure 2.2: Sparse representation classification for ATR. The input test data is represented as a linear combination of training data from a dictionary by a sparse coefficient vector [4].

(LVQ) algorithm and multi-layer perceptrons (MLPs) are used for recognition. A related architecture combining several individual ATR classifiers was proposed in [31]. A K-nearest-neighbor (KNN) data-driven approach for animal recognition using IR sensors was proposed in [102].

Recently, sparse representation-based classification (SRC) methods have shown great promise in face recognition [11, 103] and have also been applied to IR data for target detection [104], tracking and recognition [4, 105]. In [4] SRC has been shown to outperform other traditional ATR algorithms including CNN, MNN and LVQ, etc. The basic idea is to create a dictionary of training data represented as column vectors (Figure 2.2), and during this process different dimension reduction methods are used to reduce the dimension of both training and testing data. Then given a testing data, the rest is simply solving an $l_1$ minimization problem in order to find a sparse coefficient vector to represent the input data as a linear combination of the training data from the dictionary.

The main drawbacks of these data-driven approaches are that they require large sets of training data, especially in the IR ATR applications considered here, and that the profound appearance variability of the observed target thermal signatures expected under practical field conditions tends to make dictionary selection extremely difficult.

*ii. Model-Driven Approaches*

The model-driven approaches are based on computer-generated models (e.g., CAD models) with or without real-world sensor data for model learning. CAD models (Figure 2.3) have been widely used for object segmentation, tracking and recognition [5, 106, 107]. Modern model-based ATR approaches generate target hypotheses and match the observed sensor data to the hypothesized signatures or appearance models [2]. The main idea is that a better interpretation of the scene and target can be achieved by applying intelligent reasoning while preserving as much target infor-

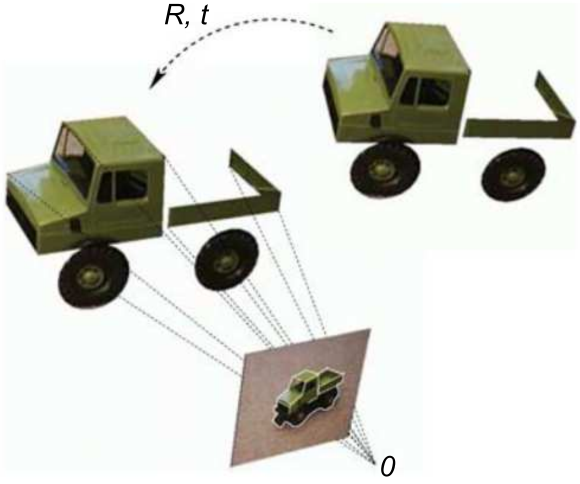


Figure 2.3: Estimate the rigid transformation that produces the object contour in the scene [5].

mation as possible [108]. For example, in [109], radar features were extracted from the sensor data and used to construct a 3D model of the observed target that was compared with known models of the objects of interest to find the best match. There are also hybrid techniques [6, 110, 111] that combine both CAD models and data-driven 2D image features for model learning and inferencing (Figure 2.4). Indeed, 2D image features play an important role in many IR ATR algorithms [112], and a variety of such shape features were evaluated in [113]. One example of a hybrid technique is the multi-view morphing algorithm that was used in [114] to construct a view morphing database in an implicit way.

A number of manifold learning methods have also been proposed for shape modeling and have been applied recently in object tracking and recognition [115]. Elliptic Fourier descriptors were used in [7] to model shapes as sums of elliptic harmonics, and the latent space of target shapes was learned through GPLVMs (Figure 2.5). In [10], a level set framework was developed to optimize a pixel-wise posterior in the shape latent space in order to achieve simultaneous segmentation and tracking. A similarity space was added in [52] to handle multi-modal problems where an efficient discrete

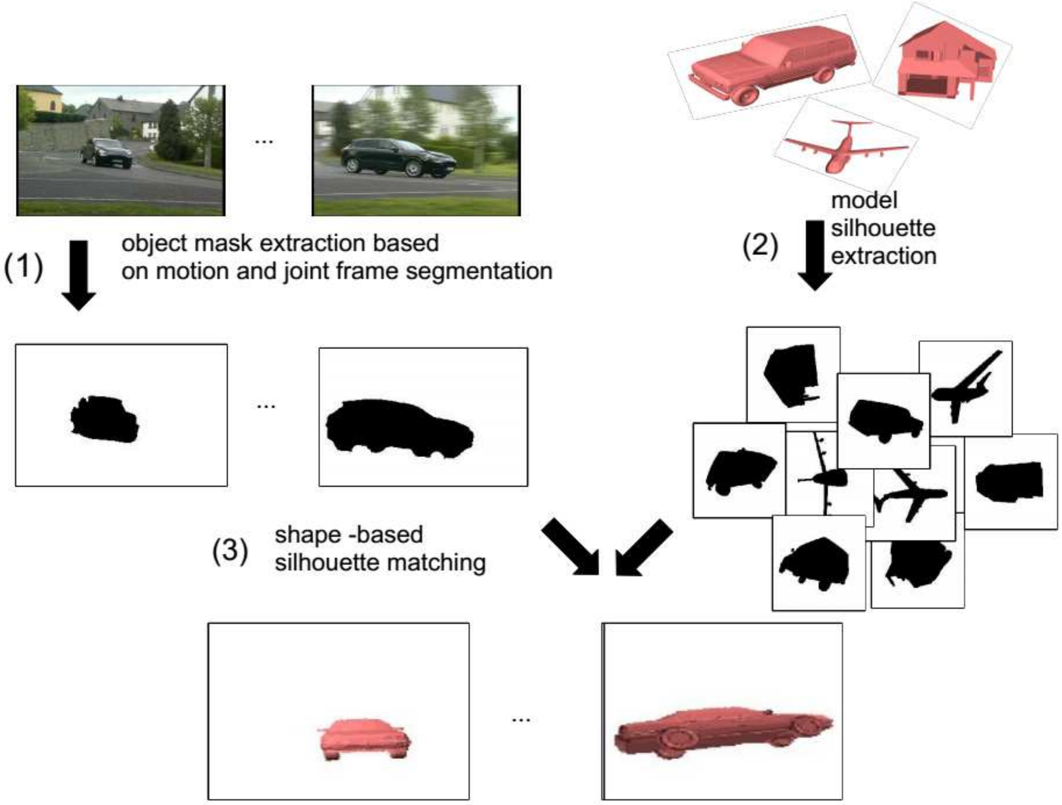


Figure 2.4: Recognition in videos by matching the shapes of object silhouettes obtained using motion segmentation with silhouettes obtained from 3D models [6].

cosine transform (DCT)-based shape descriptor was used for manifold learning. A shape model called the couplet of view and identity manifolds (CVIM) that represents the target view and identity variables on a coupled pair of manifolds was proposed in [9] for joint target tracking and recognition in IR imagery. In [53], a probabilistic level set framework with shape modeling was proposed for target tracking and recognition, where a motion model was used in a particle filter-based sequential inference process. Sampling was performed in a local area predicted by the motion model, thereby alleviating the multi-modal optimization problem.

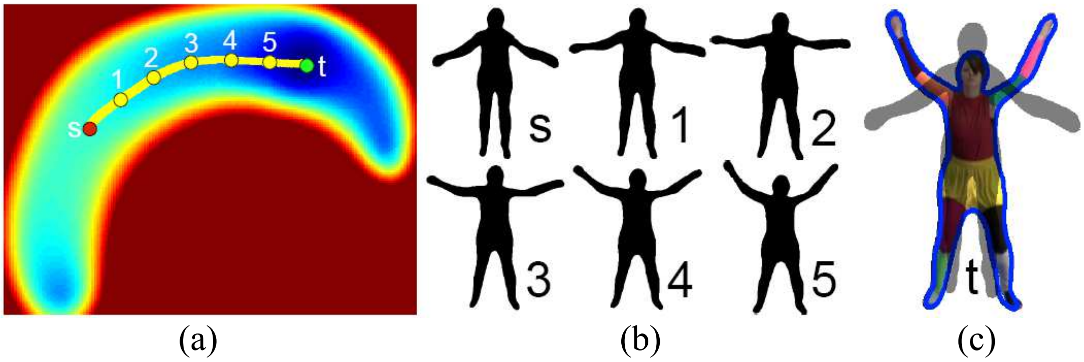


Figure 2.5: Minimization of the energy function in the latent space for segmentation. (a) values of the function in the latent space; (b) optimization trajectory starting from $s$ to 5; (c) segmentation result in blue overlayed with initialization in gray superimposed on the input image of a human [7].

## 2.2 Scene Understanding and Image Parsing

In recent years, more and more attractions have been focused on scene understanding researches, which is one of the most important application to mimic human visual systems on recognizing the world both comprehensively and accurately. One common scene understanding work is to assign single labels to given test images [116], for example suppose given a testing image the output of the algorithm is simply just an image of forests, an image of streets, or an image of a dinning room. More advanced recognition approaches are able to assign a group of labels to several objects appearing in the image [117]. In [8] a hierarchical generative model was proposed to capture not only the overall scene, but also the co-occurrences of objects and thus gives the segmentation and recognition of each object component (Figure 2.6). An interesting scene labeling system with a combination of different probabilistic classifiers with incorporated semantic context was proposed in [96].

In a more specific perspective, scene labeling is a significant part of researches on general image understanding for computer vision tasks, which tends to segment

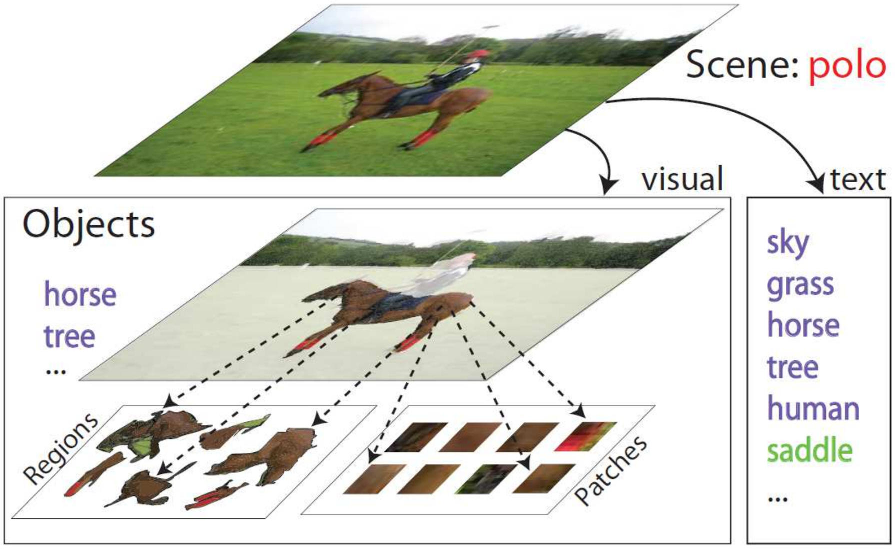


Figure 2.6: A hierarchical generative model proposed in [8] for classification, annotation and segmentation.

images into semantically meaningful region blobs with correct labels corresponding to relative classes. Traditionally, intensive studies have been focus on discriminative feature extraction from objects, and a number of low-level or mid-level human designed features are studied to statistically capture different objects in the images. But lower-level features are not statistically sufficient enough to represent discriminative feature of objects in the natural scenes, and they suffer from the problems of high dimensionality as well. [84] alleviated those issues caused from human engineered features by training a multi-scale convolutional neural network that is able to produce powerful representation of texture, shape and also contextual information, and all weights are shared to reduce the number of controllable parameters. [75] proposed a recurrent CNN that allows large input context with limited model capacity that is able to model complex spatial dependencies with relatively low inference cost, and another intra-layer recurrent connections in the convolutional layers were proposed in [118] where each convolutional layer becomes a 2D recurrent network to integrate feature extraction and context modulation. A recursive context propagation neural network was propose in [64] where a bottom-up aggregation of local information to global representation and a top-down propagation of aggregated information approach is shown to enhance the contextual information of local features. And [119] substitute conventional split nodes of classification trees with randomized multilayer perceptrons that is capable of learning possibly non-linear data representations by hidden layers.

On the other hand, graphical models have been developed to construct higher level representations by probabilistic spatial dependencies of pixels/regions among objects. For example [70] proposed a conditional random fields (CRF) based scene labeling model training with aggregated whole image features on many unlabeled nodes by marginalizing out the unknown labels. In [71] a second order sparse CRF is formulated with both unary and pairwise features to learn the conditional distribution over the class labeling, and a fully connected CRF was used in [72] which encodes spatial relationships among different objects while preserving object contours. [120] designed a CRF framework with higher order potential functions, while [74] integrated multiple levels of features by a hierarchical CRF model. But the main drawbacks of traditional learning based methods suffer from the imbalanced distribution of objects appearing in both indoor and outdoor environment, since most areas of the images are dominated by most common background classes, which makes traditional learning methods not sufficiently robust to cover all labeling situations.

Moreover, nonparametric methods involving image retrieval and label transfer have achieved promising performance on large-scale data, which try to build class likelihood of pixels/superpixels by semantically transferring labels from training images that are similar to the query image by nearest neighbors, so that irrelevant images from the training dataset might be skipped regarding to the global contextual information semantically. The most important motivation of this group of work is based on an interesting observation, which found that the scene layout of a single

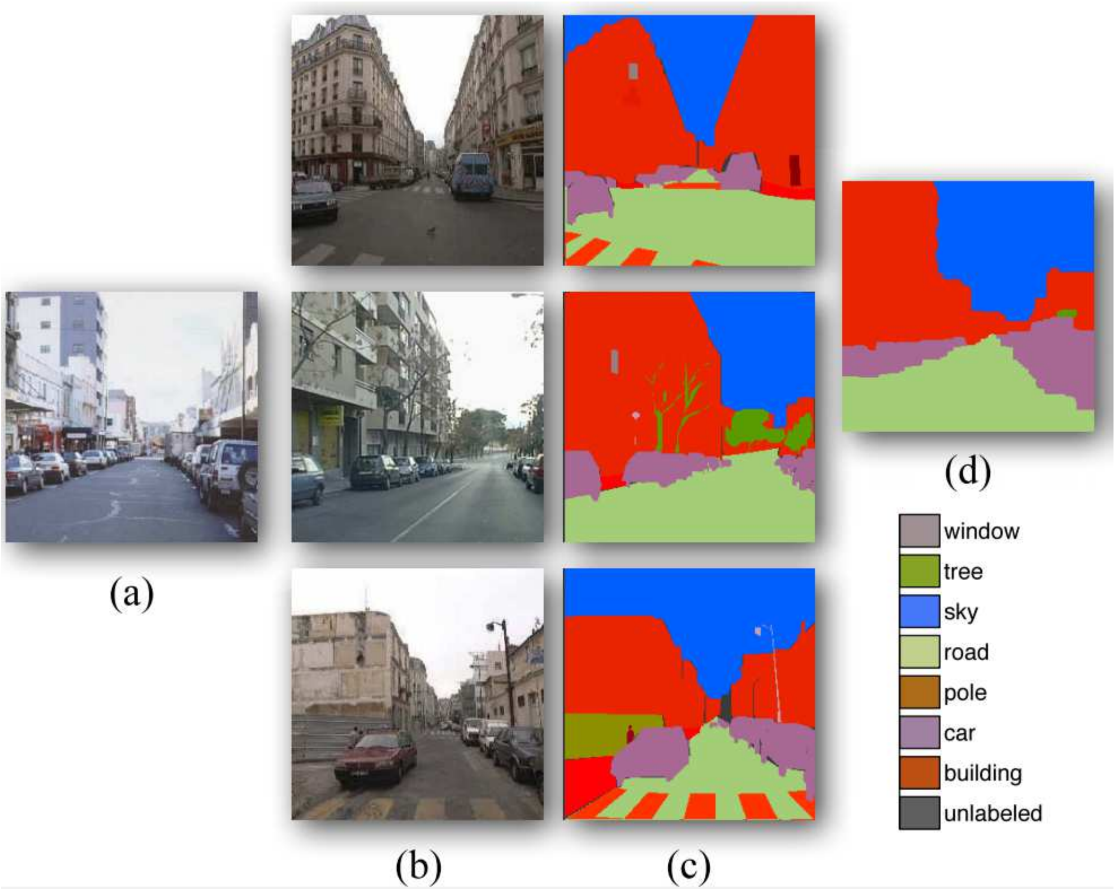


Figure 2.7: Label transfer method [1]. Top matches from the training data in (b) are found for a testing image (a), and then the known annotation (c) of top matches are transfered to label the given query image as shown in (d).

outdoor or indoor image contains only limited number of object/background categories [93]. [1] proposed a nonparametric Markov random field (MRF) framework to transfer the annotation from the best matches retrieved from a large database with annotated images to the input image (as shown in Figure 2.7) by using a combination of GIST (global meaning/structure of the scene) matching and SIFT flow algorithm. Later on the SIFT flow was extended into a hierarchical computational framework for improved performances in [85]. In [91] dense overlapping patch correspondences were constructed into a graph for mappings. [90] introduced a mid-level windows instead of low-level superpixels to capture entire objects rather than fragments. Recently [13] integrated the non-parametric label transfer method into an parametric CNN framework, which leveraged the global scene context to alleviate the local ambiguity by using learned CNN features instead of human engineered ones.

Furthermore, additional context regulation has also been addressed to incorporate global and/or neighboring information to improve semantic labeling. In [121] images are grouped into clusters and a CRF model is learned for each cluster separately. [86] initializes the labeling by pure appearance information for context extraction, and then a context index is built in the neighborhood of each superpixel. [87] proposed semantic label descriptor to refine the retrieval set obtained by global image features (GIST, color histogram and spatial pyramid over SIFT). [93] not only refined the image retrieval and superpixel matching by global semantic context on classification likelihood maps, but also expanded the retrieval set by adding rare classes exemplars for more a balanced frequency of rare classes in the retrieval set.

## 2.3 Research Motivation

Motivated by [9, 52, 53], our focus here is on developing a model-driven approach by combining relatively simple CAD models with advanced manifold learning for robust ATR to reliably segment and recognize target chips in the sequence of images acquired from an imaging IR sensor. More specifically, our goal is to incorporate CVIM into a probabilistic level set framework with shape-constrained latent space to achieve joint and seamless target segmentation, recognition and pose estimation with an analytical formulation that facilitates efficient sampling-based or gradient-based global solution of the multi-modal optimization problem (as shown in Figure 2.8). This leads to a new approach that does not require labeled training data and is free from the need of any explicit feature extraction technique. Unlike many ATR algorithms, including [9], it is also free from the need for auxiliary background rejection or pre-segmentation processing; with our proposed methods, target segmentation instead becomes a useful byproduct of the joint ATR inference process.

But due to the limitation of shape features in IR data, we explore some sparse representation based data-driven methods on a much challenging dataset as well. In this approach we incorporates the original SRC method into the multi-attribute group constraint method by an sparsity concentration index, which determines if the initial sparse coefficient vector of the original SRC method is sparse enough.

Moreover, scene labeling builds a bridge towards better scene understanding including object detection, segmentation and recognition. However, in common computer vision applications, objects of interest typically occupy only small regions in the natural scene image or FLIR data. Objects with insufficient appearing frequencies and limited coverage of the whole image intensifies the difficulty of learning a robust recognition system. We intend to build a rare class-oriented scene labeling framework which fuses parametric CNN model for local labeling and non-parametric label transfer process for global labeling [13], in which we selectively add rare class pixels to expand the retrieved image exemplar subset assisted by the scene information, rather than random sampling [93]. To further alleviate imbalanced training data problem, a complementary low-cost rare classes balanced CNN structure is trained to improve labeling performances in potential rare class regions.

In the next chapter, we will introduce a probabilistic formulation of the proposed shape manifold aware level set framework for joint target segmentation and recognition. Then we will present our MCMC-based method and two PSO methods for joint ATR optimization on the SENSIAC dataset, where targets have more preferable boundaries and less complex surrounding area/background. The first one involves a three-stage multi-threaded inference algorithm, where the solution is achieved by level set-based shape optimization and CVIM-based shape inference alternately. The second one is the standard PSO that involves CVIM-based sampling for shape interpolation, while the third one is a gradient-boosted PSO (GB-PSO) that involves a

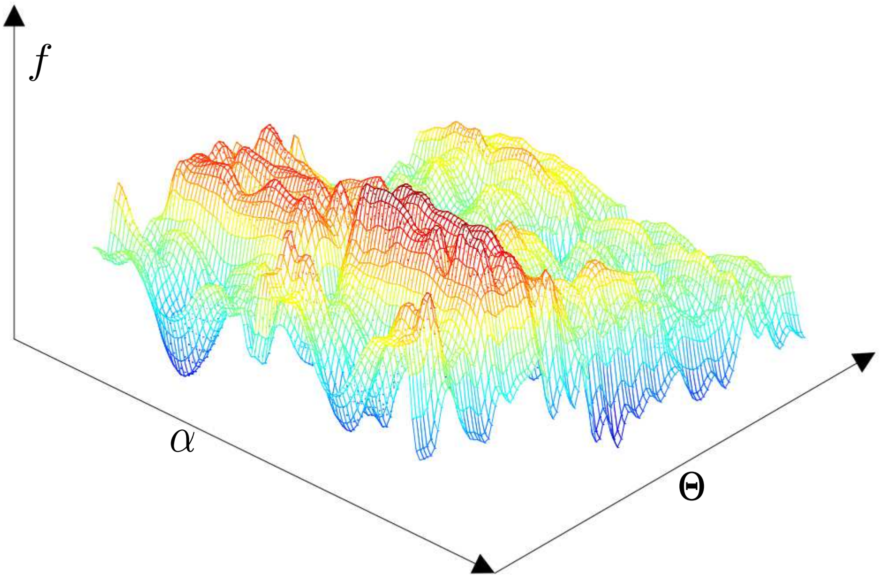


Figure 2.8: The multi-modal nature of the proposed energy function $f$ to be optimized in latent space $\{\alpha, \mathbf{\Theta}\}$.

gradient-based search in the CVIM latent space. Moreover in Chapter IV some data-driven target recognition researches using sparse representation classification methods are discussed on the Comanche dataset, where heavy background clutters and larger variation of target signatures are encountered. We will apply our shape based algorithms on the SENSIAC dataset and sparse representation based algorithms on the Comanche dataset.

# CHAPTER III

# JOINT TARGET RECOGNITION AND SEGMENTATION

In this chapter, we will first introduce some preliminary work on generative shape modeling, and a probabilistic level set graphic model, and then we will present our probabilistic formulation of the proposed shape manifold aware level set framework for joint target segmentation and recognition, followed by three different methods for joint optimization of target shape, identity and view on the SENSIAC dataset.

## 3.1 Preliminary Work

Two preliminary works are introduced in this section. The first one is the couplet view-identity manifolds (CVIM) shape model, which define a latent space as the constraint of the shape embedding function $\mathbf{\Phi}$, and the second one is the probabilistic level set framework. We will add the CVIM latent space into the level set framework in order to jointly estimate target pose and identity in the latent space defined by CVIM.

### 3.1.1 DCT-Enhanced Generative Shape Modeling

In this section, we briefly review the CVIM [9] and then extend it to accommodate DCT-based shape descriptors for learning and inference. The CVIM can be learned from a set of 2D shape silhouettes [122] created by a series of 3D CAD models by a nonlinear kernelized tensor decomposition, as shown in Figure 3.1. Here, we use six models for each of six target classes, as shown in Figure 3.2. The CVIM consists of a hemisphere-shaped view manifold and a closed-loop identity manifold in the tensor

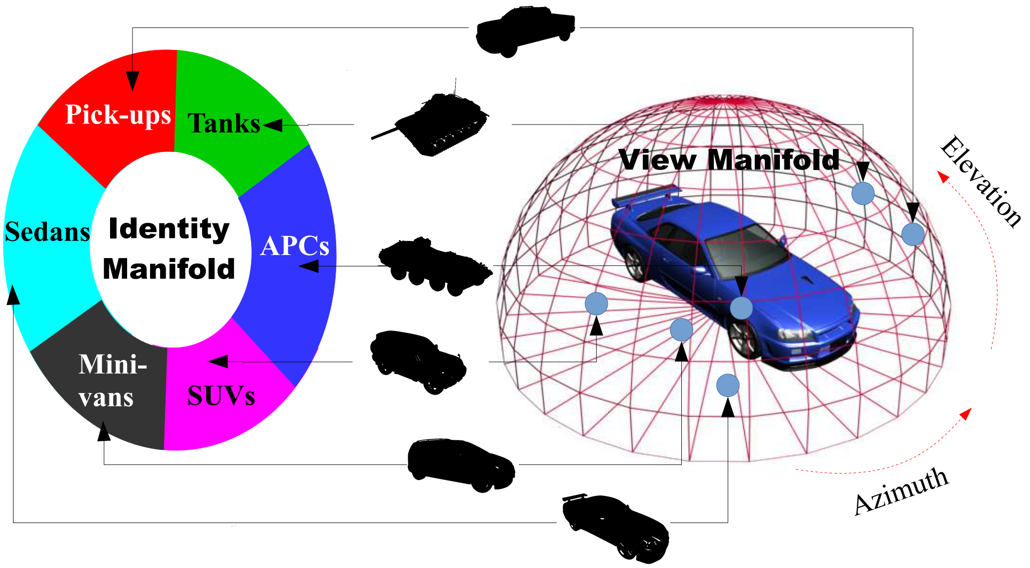


Figure 3.1: The couplet of view and identity manifolds (CVIM) shape generative model [9].

coefficient space. Two practical considerations lead to this heuristic simplification of the identity manifold. First, the SENSIAC targets of interest in this work are all man-made vehicles that exhibit distinct inter-class appearance similarities. Second, these similarities can be leveraged to judiciously order the classes along a 1D closed loop manifold in order to support convenient identity inference, as shown in Figure 3.1. In [9], a class-constrained shortest-closed-path method was proposed to deduce an optimal topology ensuring that targets of the same class or of similar shapes remain close along the identity manifold (*i.e.*, armored personnel carriers (APCs) → tanks → pick-ups → sedans → minivans → SUVs → APCs).

The original CVIM is learned as follows: (1) given a set of silhouettes (represented by the signed distance transform) from $N_m$ target types under $N_c$ views, a mapping from a conceptual hemispherical view manifold $\mathbf{\Theta}$ to the high dimensional data is learned using radial basis functions (RBFs) $\Psi(\mathbf{\Theta})$ for each target shape. (2) by stacking the collection of these mappings for all target shapes and applying the high-order

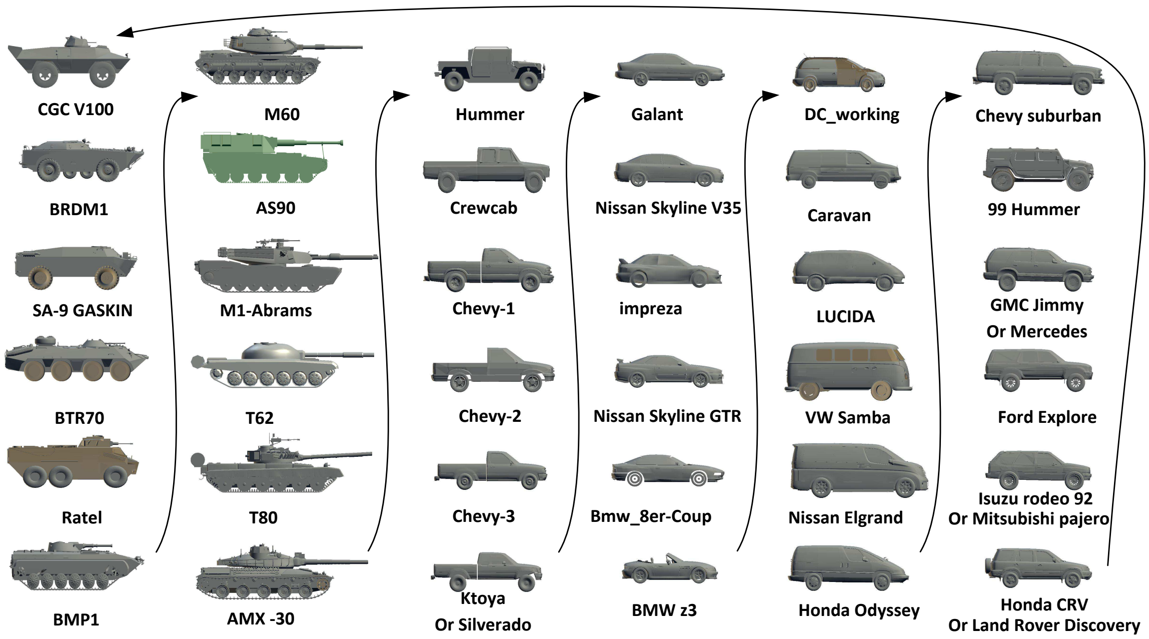


Figure 3.2: All CVIM training CAD models.

singular value decomposition (HOSVD), we obtain a core tensor $\mathbf{A}$ and $N_m$ identity vectors for all training types in the tensor coefficient space $\mathbf{i}_m$ ($m = 1, 2, ..., N_m$); (3) a mapping from the coefficient vector space to a 1D closed loop identity manifold $\alpha$ is then constructed using the optimal identity manifold topology, where each training target type $\mathbf{i}_m$ is represented by an identity vector $\mathbf{i}(\alpha_m)$ associated with a point along the identity manifold. For any arbitrary $\alpha \in [0, 2\pi)$, we can then obtain a corresponding identity vector $\mathbf{i}(\alpha)$ from the two closest training identity vectors $\mathbf{i}(\alpha_m)$ and $\mathbf{i}(\alpha_{m+1})$ by applying cubic spline interpolation along the identity manifold. The CVIM model was tested against the SENSIAC database [3] for target tracking and recognition in [9], where the experimental results validated its efficacy both qualitatively and quantitatively.

Due to the continuous nature of two manifolds, CVIM can be used to represent shapes $\mathcal{S}$ of unknown vehicles under arbitrary view point, which is especially desirable for tracking and recognition from image sequences:

$$\mathcal{S}(\alpha, \mathbf{\Theta}) = \mathbf{A} \times_3 \mathbf{i}(\alpha) \times_2 \Psi(\mathbf{\Theta}), \tag{3.1}$$

where $\mathbf{A}$ is a core tensor obtained by tensor decomposition, $\times_n$ is the mode-$n$ tensor multiplication, $\alpha$ and $\mathbf{\Theta}$ are latent variables of the identity and view on the identity and view manifold respectively, $\mathbf{i}(\alpha)$ is the identity vector obtained by cubic spline interpolation in the tensor coefficient space, and $\Psi(\mathbf{\Theta})$ is the Radial Basis Function mapping along the view manifold. Through this CVIM shape modeling, target silhouettes can be interpolated given arbitrary view point and identity variable on the 1D closed loop identity manifold. The CVIM model was tested against the SENSIAC database [3] for target tracking and recognition in [9], where the experimental results validated its efficacy both qualitatively and quantitatively.

Here, we reduce the inference complexity of the original CVIM method [9] by replacing the silhouettes used for training with the simple, but efficient DCT-based shape descriptor proposed in [52]. Thus, each training shape (in the form of the signed distance transform) is represented by a small set of 2D DCT coefficients reshaped into a column vector, where, e.g., only the top 10% of the DCT coefficients that are largest in magnitude are retained. The CVIM can then be learned using the same process as before to represent the sparse DCT coefficient vectors $\mathcal{S}_{\text{DCT}}$ of the training targets by $\mathbf{\Lambda} = [\mathbf{\Theta}^T, \alpha]^T$ according to:

$$\mathcal{S}_{\text{DCT}}(\mathbf{\Lambda}) = \mathcal{S}_{\text{DCT}}(\alpha, \mathbf{\Theta}) = \mathbf{A} \times_3 \mathbf{i}(\alpha) \times_2 \Psi(\mathbf{\Theta}) \tag{3.2}$$

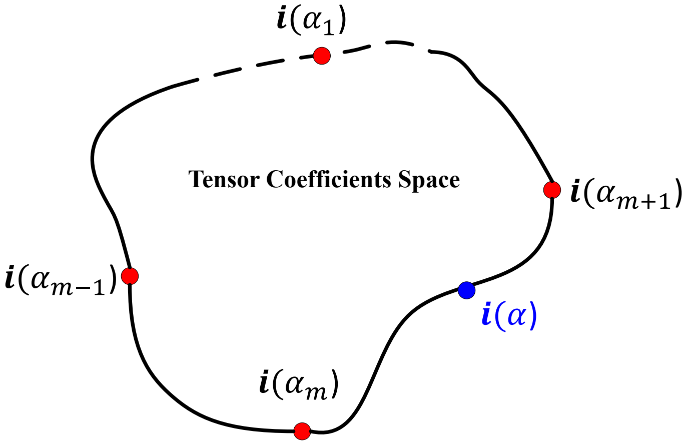
where $\mathbf{A}$ is a core tensor obtained by tensor decomposition, $\times_n$ is the mode-$n$ tensor multiplication, $\alpha$ and $\mathbf{\Theta} = [\theta, \phi]^T$ are the identity and view latent variables on the identity and view manifolds, respectively, and $\theta$ and $\phi$ are the azimuth and elevation angles. For an arbitrary $\alpha$, the associated $1 \times N_m$ identity (row) vector $\mathbf{i}(\alpha)$ in Equation (3.2) can be interpolated as:

$$\mathbf{i}(\alpha) = \mathbf{a}_m(\alpha - \alpha_m)^3 + \mathbf{b}_m(\alpha - \alpha_m)^2 + \mathbf{c}_m(\alpha - \alpha_m) + \mathbf{d}_m, \ \alpha \in [\alpha_m, \alpha_{m+1}) \tag{3.3}$$

where $\mathbf{a}_m$, $\mathbf{b}_m$, $\mathbf{c}_m$ and $\mathbf{d}_m$ are the piecewise polynomial coefficient row vectors obtained by applying cubic spline interpolation in the tensor coefficient space between the closest two adjacent training target types $\mathbf{i}(\alpha_m)$ and $\mathbf{i}(\alpha_{m+1})$, as depicted in Figure 3.3. Let $\Psi(\boldsymbol{\Theta})$ be the RBF mapping along the view manifold given by:

$$\Psi(\boldsymbol{\Theta}) = [\kappa(\|\boldsymbol{\Theta} - \mathbf{S}_1\|), ..., \kappa(\|\boldsymbol{\Theta} - \mathbf{S}_{N_c}\|)] \tag{3.4}$$

where $\kappa(\|\boldsymbol{\Theta} - \mathbf{S}_i\|) = e^{-c(\boldsymbol{\Theta}-\mathbf{S}_i)^T(\boldsymbol{\Theta}-\mathbf{S}_i)}$, $c$ is the RBF kernel width, $\mathbf{S}_i$ is a training view on the view manifold and $N_c$ is the number of training views. One major advantage of this DCT-based shape representation over the original silhouette-based one is that it naturally provides reconstruction of a shape at arbitrary magnification factors by appropriately zero-padding the DCT coefficients prior to inverse DCT (IDCT). This feature is desirable to deal with various IR targets at different ranges.



$\boldsymbol{i}(\alpha_1)$
Tensor Coefficients Space
$\boldsymbol{i}(\alpha_{m+1})$
$\boldsymbol{i}(\alpha_{m-1})$
$\boldsymbol{i}(\alpha)$
$\boldsymbol{i}(\alpha_m)$

Figure 3.3: Cubic spline interpolation along the identity manifold in CVIM.

We represent the shape embedding function $\boldsymbol{\Phi}$ (referred to in Figure 3.5c and Equation (3.15)) in terms of the CVIM parameter $\boldsymbol{\Lambda}$ by:

$$\boldsymbol{\Phi}(\boldsymbol{\Lambda}) = \mathsf{IDCT}(\mathcal{S}_{\text{DCT}}(\boldsymbol{\Lambda})) \tag{3.5}$$

where $\mathsf{IDCT}(\cdot)$ is the IDCT with two reshape operations. The first is for the input (from 1D to 2D) prior to IDCT, and the second is for the output (from 2D to 1D) after IDCT. Note that the derivative of the IDCT of a matrix may be computed as the IDCT of the derivative of that matrix [52]. Therefore, the DCT-shape presentation can easily be incorporated into the above optimization framework without major modifications. Through this CVIM model, target shapes corresponding to arbitrary $\alpha$ can readily be interpolated along the view and identity manifolds.

### 3.1.2 Pixel-Wise Posterior Level Set Segmentation

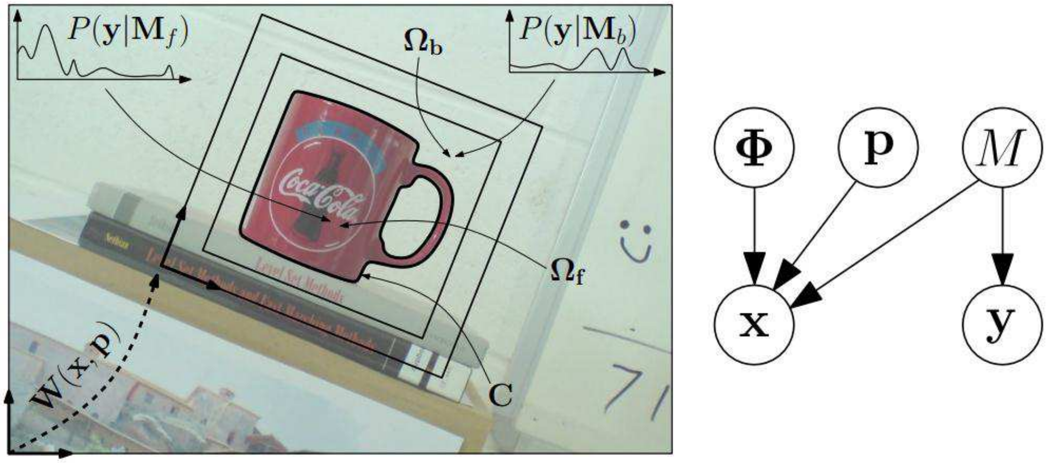


Figure 3.4: Left: Representation of a object: the contour $C$, foreground $\Omega_f$ and background $\Omega_b$, foreground/background models $M$, and the warp $W(\mathbf{x}, \mathbf{p})$. Right: graphical representation of a probabilistic level-set framework, where $\mathbf{p}$ is the parameter of a warp function $W$, $\mathbf{x}$ is a pixel location, $\mathbf{y}$ is a pixel value. [10]

The robustness of using an implicit contour or level-set to represent boundary of an object has been proved in recent years. In [10] a probabilistic level-set framework (Figure 3.4) was proposed, which gives a probabilistic interpretation of most region based level-set methods, and by introducing a pixel-wise posterior term as a energy function, model parameters are marginalised out at a pixel level, then segmentation can be easily achieved through variational level set evolution.

From the graphical model in Figure3.4, a joint distribution is:

$$P(\mathbf{x}, \mathbf{y}, \mathbf{\Phi}, \mathbf{p}, M) = P(\mathbf{x}|\mathbf{\Phi}, \mathbf{p}, M)P(\mathbf{y}|M)P(\mathbf{\Phi})P(\mathbf{p})P(M), \tag{3.6}$$

divide (3.6) by $P(\mathbf{y})$, then marginalising over the model M, and by using a logarithmic opinion pool (LogOP), a pixel-wise posterior was derived:

$$P(\mathbf{\Phi}, \mathbf{p}|\Omega) = \prod_{i=1}^{N} \Big\{ \big(P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}_i)\big) \Big\} P(\mathbf{\Phi})P(\mathbf{p}), \tag{3.7}$$

where $N$ is number of pixels, and $i$ is pixel index. $P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}_i)$ is defined as:

$$P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}_i) = H_\epsilon(\mathbf{\Phi}(\mathbf{x}_i))P_f + (1 - H_\epsilon(\mathbf{\Phi}(\mathbf{x}_i)))P_b, \tag{3.8}$$

where

$$P_f = \frac{P(\mathbf{y}_i|M_f)}{\eta_f P(\mathbf{y}_i|M_f) + \eta_b P(\mathbf{y}_i|M_b)}, \tag{3.9}$$

$$P_b = \frac{P(\mathbf{y}_i|M_b)}{\eta_f P(\mathbf{y}_i|M_f) + \eta_b P(\mathbf{y}_i|M_b)}, \tag{3.10}$$

where $\eta_f$ and $\eta_b$ are number of pixels belong to the foreground and background region respectively, $P(\mathbf{y}_i|M_f)$ and $P(\mathbf{y}_i|M_b)$ are foreground and background models represented by histograms, and $H_\epsilon(\cdot)$ is the smoothed Heaviside step function. Here the prior of the shape embedding function $P(\mathbf{\Phi})$ that encourages $\mathbf{\Phi}$ to resemble a signed distance function as:

$$P(\mathbf{\Phi}) = \prod_{i=1}^{N} \big[\frac{1}{\sigma\sqrt{2\pi}} \exp\big(-\frac{(|\nabla\mathbf{\Phi}(\mathbf{x}_i)| - 1)^2}{2\sigma^2}\big)\big], \tag{3.11}$$

where $\sigma$ gives the relative weight of the prior. Substitute (3.8) and (3.11) into (3.7) and take log gives the final expression of the log posterior:

$$\begin{aligned}\log(P(\mathbf{\Phi}, \mathbf{p}|\Omega)) \propto \sum_{i=1}^{N} \Big\{ \big(\log(P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}_i)) - \frac{(|\nabla\mathbf{\Phi}(\mathbf{x}_i)| - 1)^2}{2\sigma^2}\big) \Big\} \\ + N\log(\frac{1}{\sigma\sqrt{2\pi}}) + \log(P(\mathbf{p})).\end{aligned} \tag{3.12}$$

For segmentation purposes, the registration $\mathbf{p}$ is assumed to be known, then by taking the derivative of (3.12) with respect to $\mathbf{\Phi}$, and then by gradient flow [10] a shape embedding function $\Phi$ that maximize the final log posterior (3.12) will be found. This probabilistic level set framework define us an energy function as pixel-wise posterior, and as mentioned in Chapter II, if a shape model was trained through dimension reduction methods by which the level set embedding function can be constrained in a latent space, a so-called model-based target recognition framework can be extended from this basic level set energy function.

## 3.2 Problem Formulation

Implicit contour and level set methods have been proven effective for image segmentation by optimizing an energy function, which represents the contour of an object appearing in the scene. A common approach is to compute the segmentation by optimizing the shape embedding function $\mathbf{\Phi}$ [123]. The basic idea is to initialize a shape contour and then minimize the energy function related to $\mathbf{\Phi}$ along the gradient direction. A probabilistic level set segmentation framework was proposed in [10], where, as illustrated in Figure 3.5a, an energy function called the pixel-wise posterior was defined to represent the image as a bag of pixels with the background and foreground models obtained from $\mathbf{\Phi}$ [124].

Here, we extend the model from [10] to obtain a new shape-constrained level set segmentation method by incorporating the CVIM shape model parameterized by $\mathbf{\Lambda} = [\mathbf{\Theta}^T, \alpha]^T$, which explicitly represents the target identity variable $\alpha$ and azimuth/elevation view angles $\mathbf{\Theta} = [\theta, \phi]^T$, thus inherently supporting joint target recognition, segmentation and pose estimation. We derive a new joint probability density function:

$$P(\mathbf{x}, \mathbf{y}, \mathbf{\Lambda}, \mathbf{p}, M) = P(\mathbf{x}|\mathbf{\Lambda}, \mathbf{p}, M)P(\mathbf{y}|M)P(M)P(\mathbf{\Lambda})P(\mathbf{p}) \tag{3.13}$$

where $M$ is the foreground/background model, $\mathbf{p}$ is the location of the target centroid in image coordinates, $P(\mathbf{p})$ is the prior probability of the target centroid location,

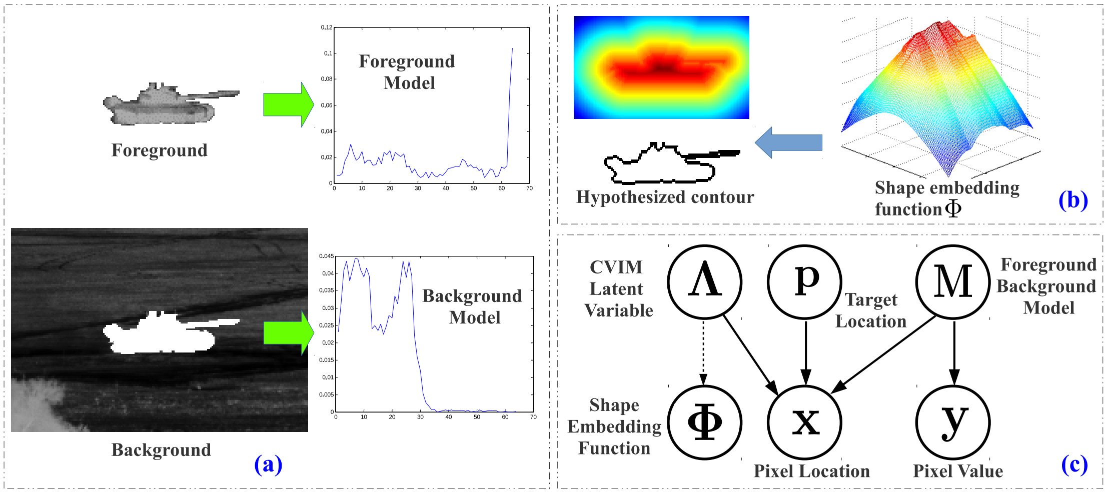


Figure 3.5: (**a**) Representation of an IR target by a hypothesized shape contour that separates the foreground and background regions; (**b**) shape embedding function $\mathbf{\Phi}$ represented by the signed distance transform. $\mathbf{\Phi}$ is generated by CVIM given the parameter vector $\mathbf{\Lambda}$, which contains the view angles $\mathbf{\Theta}$ and identity variable $\boldsymbol{\alpha}$. (**c**) The proposed probabilistic level set framework, where $\mathbf{p}$ is the target centroid in image coordinates, $M$ is the foreground/background model, $\mathbf{x}$ is a pixel location in image coordinates and $\mathbf{y}$ is a pixel intensity value. The dashed line represents the CVIM-based mapping from the latent shape space $\mathbf{\Lambda}$ to the shape embedding function $\mathbf{\Phi}$.

which is assumed uniform. $\mathbf{x}$ is a pixel location in image coordinates and $\mathbf{y}$ is the pixel intensity. The intensity is usually scalar-valued for the case of an imaging MWIR or LWIR sensor, but may generally be vector-valued in our framework and formulation.

By marginalizing over the foreground/background model $M$ [10] and using the logarithmic opinion pool [125], we formulate a new pixel-wise posterior:

$$P(\mathbf{\Lambda}, \mathbf{p}|\Omega) = \prod_{i=1}^{N} \left\{ P(\mathbf{x}_i|\mathbf{\Lambda}, \mathbf{p}, \mathbf{y}_i) \right\} \cdot P(\mathbf{\Lambda})P(\mathbf{p}) \tag{3.14}$$

where $\Omega = \{\mathbf{x}, \mathbf{y}\}$ is a small IR chip cropped from the IR frame acquired by the sensor, $N$ is the number of pixels in the chip and $i$ is the pixel index. Because we are focused on small IR chips that contain a target, we localize the target centroid

$\mathbf{p}$ after segmentation and recognition. Therefore, $\mathbf{p}$ is omitted in the following. As in [10], $P(\mathbf{x}_i|\mathbf{\Lambda}, \mathbf{y}_i)$ in Equation (3.14) may be expressed according to:

$$P(\mathbf{x}_i|\mathbf{\Lambda}, \mathbf{y}_i) = H_\epsilon(\mathbf{\Phi}_{\mathbf{x}_i})P_f + (1 - H_\epsilon(\mathbf{\Phi}_{\mathbf{x}_i}))P_b \tag{3.15}$$

where $\mathbf{\Phi}$ is the shape embedding function generated from CVIM given $\mathbf{\Lambda}$ (in the form of a signed distance function, as shown in Figure 3.5b), $H_\epsilon(\cdot)$ is the smoothed Heaviside step function and $\mathbf{\Phi}_{\mathbf{x}_i}$ is the value of $\mathbf{\Phi}$ at pixel location $\mathbf{x}_i$. In Equation (3.15),

$$P_f = \frac{P(\mathbf{y}_i|M_f)}{\eta_f P(\mathbf{y}_i|M_f) + \eta_b P(\mathbf{y}_i|M_b)} \tag{3.16}$$

and

$$P_b = \frac{P(\mathbf{y}_i|M_b)}{\eta_f P(\mathbf{y}_i|M_f) + \eta_b P(\mathbf{y}_i|M_b)} \tag{3.17}$$

where $\eta_f$ and $\eta_b$ are the number of pixels belonging to the foreground and background regions respectively and where $P(\mathbf{y}|M_f)$ and $P(\mathbf{y}|M_b)$ are the foreground and background appearance models, which are represented by histograms.

The goal of the shape-constrained level set optimization is then to maximize Equation  (3.14) with respect to $\mathbf{\Lambda}$ according to:

$$\mathbf{\Lambda}^* = \arg\max_{\mathbf{\Lambda}} P(\mathbf{\Lambda}|\Omega) \tag{3.18}$$

The calculus of variations could be applied to compute the derivative of Equation (3.14) with respect to $\mathbf{\Lambda}$. However, due to the multi-modal nature of the CVIM-based shape modeling, we develop a PSO-based optimization framework to search for the optimal latent variable $\mathbf{\Lambda}^*$ that maximizes Equation (3.14). To further enhance the efficiency, we then develop a gradient-boosted PSO (GB-PSO) method that provides faster optimization by taking advantage of the parametric nature of CVIM.

Optimizing the posterior Equation (3.14) may be thought of as finding a contour that maximizes the histogram difference between the foreground and background in the region of interest. This consideration is based on the assumption that the target-of-interest has different intensity values compared with the background. Intuitively,

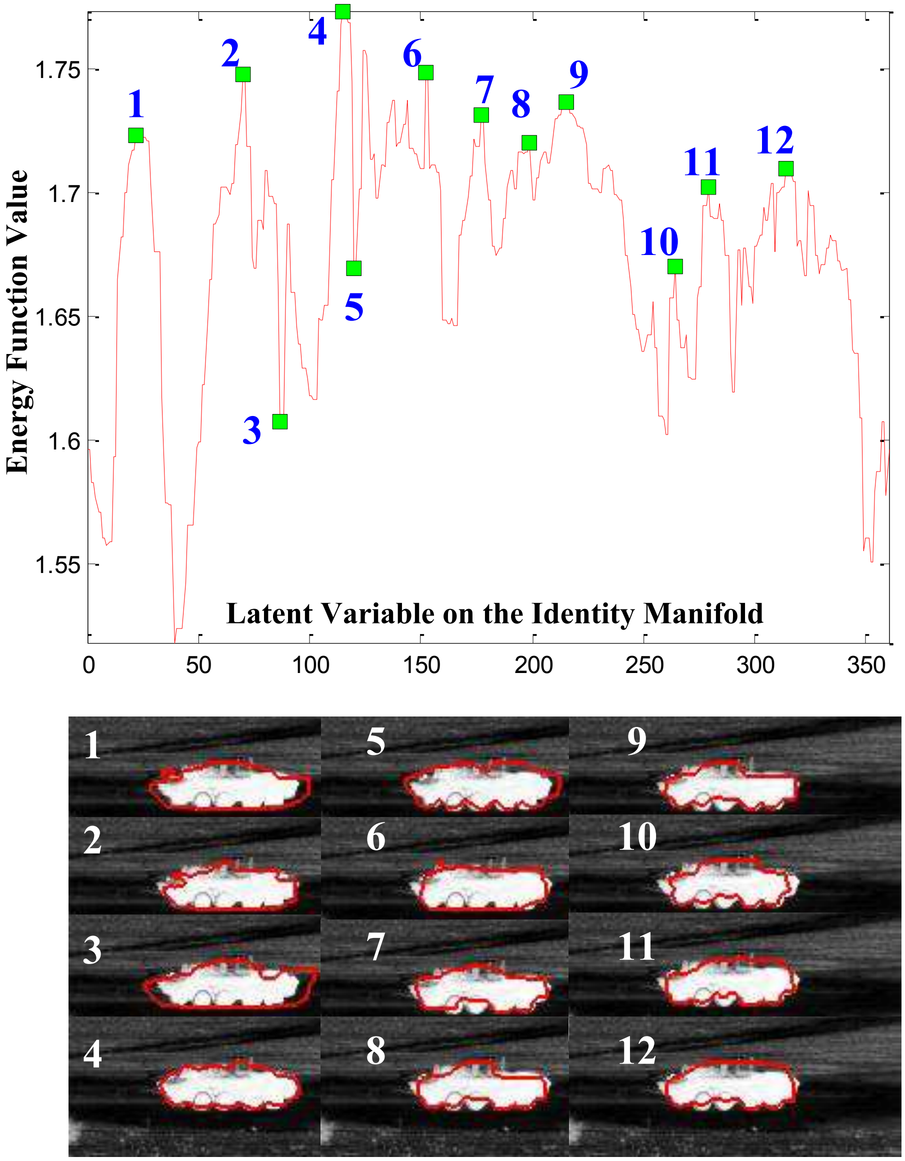


Figure 3.6: Effectiveness of the energy function. (Top) Plot of the energy function with respect to the identity latent variable; (Bottom) a group of CVIM-generated shapes corresponding to the latent variables labeled in the plot on the top are superimposed onto the original IR data chips.

if the shape contour of the target is correctly hypothesized in terms of the target type (recognition), view angle (pose estimation) and location (segmentation), then the foreground and background defined by this contour will have maximum histogram divergence and, therefore, maximize the energy function Equation (3.14) as illustrated in Figure 3.6. For a given observation, in Figure 3.6, we calculate the value of the energy function with respect to almost all possible values along the circularly-shaped identity manifold $\alpha = 1, 2, 3, ..., 360°$ with the view angle $\mathbf{\Theta}$ known for simplicity. The figure shows several CVIM interpolated shapes superimposed on the original mid-wave IR image data. As seen in the left part of the figure, the maximum value of the energy function is attained by the contour (numbered 4) that is best in the sense of being closest to the actual boundary of the target in the right part of the figure. However, the multi-model nature of the energy function as shown in Figure 3.6 (left part), which is typical, represents significant challenges for CVIM-based shape optimization and motivates the PSO and GB-PSO algorithms that we develop below in Sections 3.4 and 3.5.

## 3.3 Markov Chain Monte Carlo Optimization Method

Due to the co-existence of $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ as well as the nonlinear and multi-modal nature of the CVIM, we develop a multi-threaded optimization framework so that the energy function can be solved in parallel for each thread. Our first approach is to interleave the optimization of the target shape $\mathbf{\Phi}$ and latent variable $\mathbf{\Lambda}$ by a MCMC-based method where the CVIM is used for shape interpolation/matching (Figure 3.7). In the later chapter as better alternatives, a Particle Swarm Optimization (PSO) method and a gradient-boosted PSO method (GB-PSO) are implemented to alleviate the computational cost for interleaved multi-modal optimization and explicit shape interpolation and matching.

*(i) Multi-Threaded Optimization: Initialization*

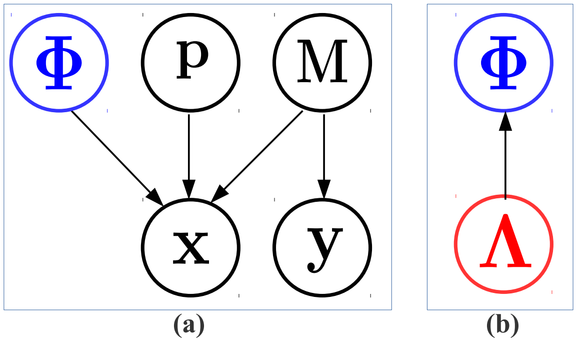


Figure 3.7: Interleaved optimization. (a) Shape inference. (b) latent space inference.

In this work, we propose a multi-threaded optimization algorithm to solve (3.18), as shown in Figure 3.8. The initialization stage has three steps to initialize a 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 ($\mathbf{\Lambda}_1^{(1:L)}$) are selected as the seeds to start the multi-threaded optimization to estimate $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ iteratively.

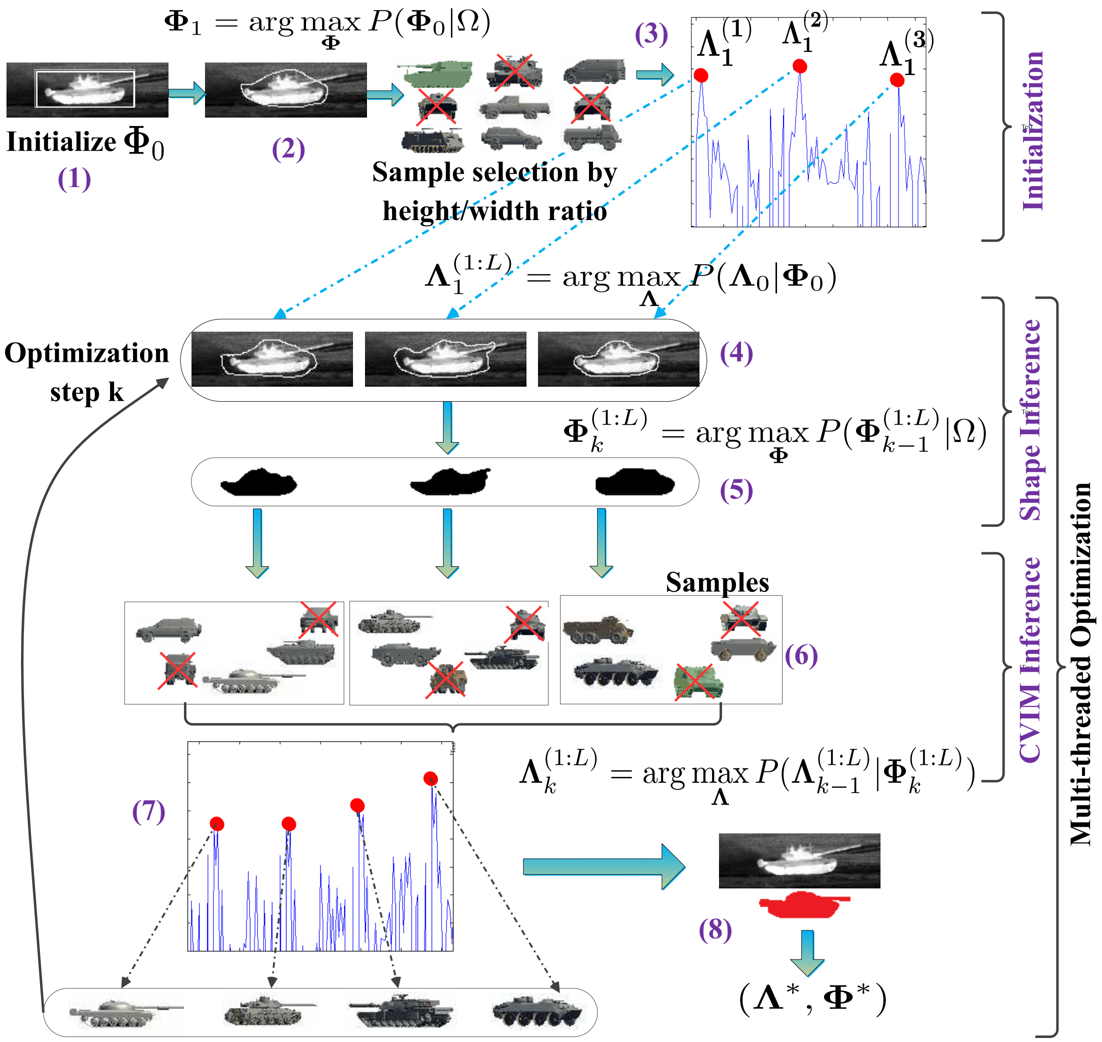
*(ii) MCMC-Based Method*

As mentioned earlier, we designed an interleaved optimization framework to optimize the latent variable $\mathbf{\Lambda}$ and the shape $\mathbf{\Phi}$ iteratively, so in this section we will introduce the shape inference and latent space inference respectively as shown in Figure 3.8.

**Shape Inference:** at this stage (Figure 3.7a), we are only looking for a shape contour that maximize the energy function (3.14) under a shape prior as

$$\mathbf{\Phi}_k^{(l)} = \arg\max_{\mathbf{\Phi}} P(\mathbf{\Phi}_{k-1}^{(l)}|\Omega), \tag{3.19}$$

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 the CVIM that is used to initialize the level set optimization.



$\boldsymbol{\Phi}_1 = \arg\max_{\boldsymbol{\Phi}} P(\boldsymbol{\Phi}_0|\Omega)$
Initialize $\Phi_0$
(1)
(2)
Sample selection by height/width ratio
(3)
$\boldsymbol{\Lambda}_1^{(1)}$
$\boldsymbol{\Lambda}_1^{(2)}$
$\boldsymbol{\Lambda}_1^{(3)}$
Initialization
$\boldsymbol{\Lambda}_1^{(1:L)} = \arg\max_{\boldsymbol{\Lambda}} P(\boldsymbol{\Lambda}_0|\boldsymbol{\Phi}_0)$
Optimization step k
(4)
$\boldsymbol{\Phi}_k^{(1:L)} = \arg\max_{\boldsymbol{\Phi}} P(\boldsymbol{\Phi}_{k-1}^{(1:L)}|\Omega)$
Shape Inference
(5)
Samples
(6)
CVIM Inference
Multi-threaded Optimization
$\boldsymbol{\Lambda}_k^{(1:L)} = \arg\max_{\boldsymbol{\Lambda}} P(\boldsymbol{\Lambda}_{k-1}^{(1:L)}|\boldsymbol{\Phi}_k^{(1:L)})$
(7)
(8)
$(\boldsymbol{\Lambda}^*, \boldsymbol{\Phi}^*)$

Figure 3.8: (1)Shape initialization $\boldsymbol{\Phi}_0$; (2)obtain $\boldsymbol{\Phi}_1$ by level set segmentation; (3)training data ($\boldsymbol{\Lambda}_0$) selection based on $\boldsymbol{\Phi}_1$, followed by the optimization of the latent variable $\boldsymbol{\Lambda}_1^{(1:L)}$, (4) start the multi-threaded optimization; (5)shape inference -for each thread $1toL$, optimize shape $\boldsymbol{\Phi}_k^{(1:L)}$; (6) after sample selection in the latent space, optimize $\boldsymbol{\Lambda}_k^{(1:L)}$; (7) Pick up local maximum in each thread for next step optimization; (8) when the optimization converge, pick up the one that has the largest energy value from step (7) as final estimation result.

Since the latent space $\mathbf{\Lambda}$ is not considered here, we can define

$$P(\mathbf{\Phi}) = \prod_{i=1}^{N} [\frac{1}{\sigma\sqrt{2\pi}} \exp\big(-\frac{(|\nabla\mathbf{\Phi}(\mathbf{x}_i)| - 1)^2}{2\sigma^2}\big)], \tag{3.20}$$

where $\sigma$ gives the relative weight of the prior. Here it encourages $\mathbf{\Phi}$ to resemble a signed distance function [10]. Substitute Equations (3.15) and (3.20) into Equation (3.14) and hence into Equation (3.19), take the log and then take the first variation with respect to $\mathbf{\Phi}$, the term $P(\mathbf{\Lambda})$ will be dropped, and here we get:

$$\frac{\partial f}{\partial \mathbf{\Phi}} = \frac{\delta_\epsilon(\mathbf{\Phi})(P_f - P_b)}{P(\mathbf{x}|\mathbf{\Phi}, \mathbf{y})} - \frac{1}{\sigma^2}[\nabla^2\mathbf{\Phi} - div(\frac{\nabla\mathbf{\Phi}}{|\nabla\mathbf{\Phi}|})], \tag{3.21}$$

where $\sigma^2 = 50$, $f = \log P(\mathbf{\Lambda}|\Omega)$, $\nabla^2$ is the Laplacian operator and $\delta_\epsilon(\mathbf{\Phi})$ is derivative of a blurred Heaviside step function, and $div(\cdot)$ is the divergence operator [122]. 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 $\tau$ is the time step [10].

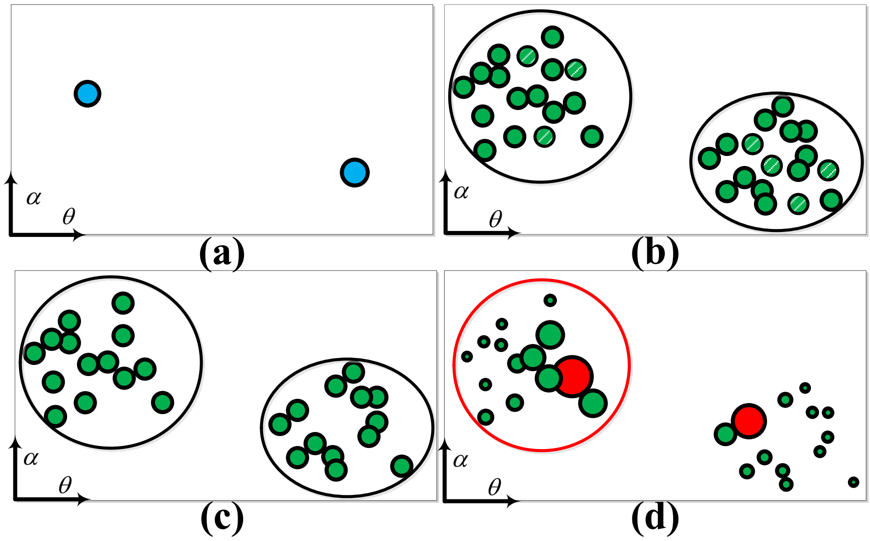


Figure 3.9: (a) The seeds in the latent space for multi-threaded shape estimation. (b) MCMC sampling for each thread. (c) Sample screening based on the height/width ratio. (d) Sample weighting by shape matching and multi-thread reset.

**Algorithm 1** MCMC-based multi-threaded optimization

```
Initialization
• Initialize a bounding box $\mathbf{\Phi}_0$ around the object
• Optimize Equation (3.19) 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, $\mathbf{\Lambda}_0^{(1:L)}$
  (Figure 3.9a)
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 Equation (3.19) to get $\mathbf{\Phi}_k^{(l)}$
    CVIM Inference
    • Optimize Equation (3.22) by MCMC
    • Draw samples around $\mathbf{\Lambda}_{k-1}^{(l)}$ in the shape space (Figure 3.9b)
    • Discard samples according to $\gamma(\mathbf{\Phi}_k^{(l)})$ (Figure 3.9c)
    • Evaluate the left samples by template matching Equation (3.23) (Figure 3.9d)
    • 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 (3.14) and $\mathbf{\Phi}^* = \mathcal{S}(\mathbf{\Lambda}^*)$ is the
  final segmentation result.
```

**Latent Space Inference:** given the shape embedding function $\mathbf{\Phi}_k$ (where we have dropped the thread index for simplicity), we will optimize $\mathbf{\Lambda}_k$ (Figure 3.7b) by performing CVIM inference as

$$\mathbf{\Lambda}_k = \arg\max_{\mathbf{\Lambda}} \{P(\mathbf{\Phi}_k|\mathbf{\Lambda}_{k-1})P(\mathbf{\Lambda}_{k-1})\}, \tag{3.22}$$

where the likelihood $P(\mathbf{\Phi}|\mathbf{\Lambda})$ is defined as a template matching:

$$P(\mathbf{\Phi}_k|\mathbf{\Lambda}_{k-1}) \propto \exp\big(-\frac{\|\mathcal{C}(\mathbf{\Phi}_k) - \mathcal{S}(\mathbf{\Lambda}_{k-1})\|^2}{2\xi^2}\big), \tag{3.23}$$

where $\mathcal{C}(\mathbf{\Phi}_k)$ is a shape silhouette obtained by thresholding $\mathbf{\Phi}_k$, $\mathcal{S}(\mathbf{\Lambda}_{k-1})$ is the CVIM-based shape interpolation given $\mathbf{\Lambda}_{k-1}$, $\|\cdot\|$ represents the shape matching error, and $\xi$ control the sensitivity of shape matching. $P(\mathbf{\Lambda}_{k-1})$ is the prior probability from previous CVIM inference. Furthermore, we developed a Markov chain Monte Carlo (MCMC)-based inference algorithm for multi-threaded CVIM inference to optimize Equation (3.22), which is interleaved with the level set shape optimization defined in Equation (3.19) iteratively. Figure 3.9 illustrates the major steps in the MCMC-based CVIM inference. We summarize the multi-threaded optimization in Algorithm 1 that combines the three stages together.

## 3.4 Particle Swarm Optimization Method

Particle swarm optimization was introduced by J. Kennedy et al. [54] for optimization of continuous nonlinear functions. The PSO algorithms were originally inspired by the observation of bird blocking and fish schooling. Unlike other genetic algorithms which are motivated by the survival of fittest - natural selection rule, the PSO is generally a simulation of social behaviors. So far, the PSO has been widely used not only in scientific researches [126–128], but also in many engineering applications [129], such as antennas [130–132], biomedical [133,134], transportation network design [135], classification [136, 137], control [138, 139], financial [140], neural networks [141–143], power systems [144], robotics [145], and signal processing [146, 147], etc.

We first implement a standard PSO algorithm due to its simplicity and effectiveness in dealing with multi-modal optimization problems. PSO optimizes a problem by moving solution hypotheses around in the search-space according to the current hypothesis and velocity computed to the present local and global optima. Our energy function is defined in Equation (3.14). Since we assume $\mathbf{\Lambda} = [\theta, \phi, \alpha]^T$ (CVIM parameters) to be uniformly distributed (*i.e.* no prior knowledge) with the registration

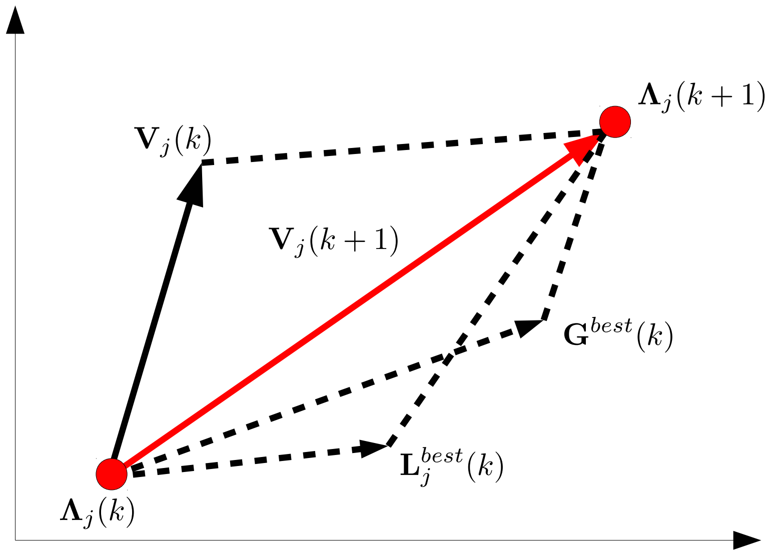


Figure 3.10: PSO update process for particle $j$. $\mathbf{\Lambda}_j(k)$ is the latent variable estimated at step $k$, $\mathbf{L}_j^{best}(k)$ and $\mathbf{G}^{best}(k)$ are local best position of particle $j$ and global best estimation in the latent space at step $k$. While $\mathbf{V}_j(k)$ is the velocity estimated at step $k$, and $\mathbf{V}_j(k+1)$ is the current updated velocity computed from Equation (3.26), and then $\mathbf{\Lambda}_j(k+1)$ is the step $k+1$ estimation of particle $j$ by Equation (3.25).

of the object frame $\mathbf{p}$ omitted, the energy function Equation (3.14) is rewritten as:

$$f(\mathbf{\Lambda}) = P(\mathbf{\Lambda}|\Omega) \propto \prod_{i=1}^{N} \left\{ H_\epsilon(\mathbf{\Phi}_{\mathbf{x}_i}(\mathbf{\Lambda}))P_f + (1 - H_\epsilon(\mathbf{\Phi}_{\mathbf{x}_i}(\mathbf{\Lambda})))P_b \right\} \tag{3.24}$$

where $\mathbf{\Phi}_{\mathbf{x}_i}(\mathbf{\Lambda})$ is defined as the value of shape embedding (in the form of the signed distance transform) in Equation (3.5) at pixel location $\mathbf{x}_i$. During the PSO optimization process, particles are updated as flying in the latent space of CVIM, $\mathbf{\Lambda}$, based on the velocity $\mathbf{V}$:

$$\mathbf{\Lambda}_j(k+1) = \mathbf{\Lambda}_j(k) + \mathbf{V}_j(k+1) \tag{3.25}$$

where $j$ and $k$ are the particle and iteration indexes, respectively. Velocity $\mathbf{V}$ is a randomly-weighted average of the best position evaluated by that particle so far and the global best position among all particles:

$$\mathbf{V}_j(k+1) = \mathbf{V}_j(k) + \boldsymbol{\Upsilon}_1 \cdot (\mathbf{L}_j^{best}(k) - \boldsymbol{\Lambda}_j(k) + \boldsymbol{\Upsilon}_2 \cdot (\mathbf{G}^{best}(k) - \boldsymbol{\Lambda}_j(k)) \tag{3.26}$$

where $\mathbf{V}_j(k)$ is the velocity for particle $j = 1 : ps$ at optimization step $k = 1 : K$ and $ps$ is the population size. $\boldsymbol{\Upsilon}_1$ and $\boldsymbol{\Upsilon}_2$ are random vectors, where each entry is uniformly distributed between $[0, 1]$. $\mathbf{L}_j^{best}(k)$ is the best position in the latent space found by particle $j$ evaluated by Equation (3.24), while $\mathbf{G}^{best}(k)$ is the global best position found among all particles.

**Algorithm 2** PSO method.

```
Initialization
• do level-set segmentation to initialize the target location $\mathbf{p}$
• draw particles $\boldsymbol{\Lambda}_j(0), j = 1 : ps$ randomly distributed in the latent space $\boldsymbol{\Lambda}$, where $ps$ is the population size
• evaluate $\boldsymbol{\Lambda}_j(0)$ by Equation (3.24) to get $f(\boldsymbol{\Lambda}_j(0))$, set $\mathbf{G}^{best}(0) = \arg\max_{\boldsymbol{\Lambda}_j(0)} f(\boldsymbol{\Lambda}_j(0))$ and $\mathbf{L}_j^{best}(0) = \boldsymbol{\Lambda}_j(0)$, $(j = 1 : ps)$
PSO algorithm
for each iteration $(k = 0 : K - 1)$ do
  for each particle $(j = 1 : ps)$ do
    • calculate velocity $\mathbf{V}_j(k+1)$ and new particle $\boldsymbol{\Lambda}_j(k+1)$ by Equations (3.26) and (3.25)
    • compute $f(\boldsymbol{\Lambda}_j(k+1))$ by Equation (3.24)
    if $f(\boldsymbol{\Lambda}_j(k+1)) > f(\mathbf{L}_j^{best}(k))$ then
      • set $\mathbf{L}_j^{best}(k+1) = \boldsymbol{\Lambda}_j(k+1))$
      if $f(\boldsymbol{\Lambda}_j(k+1)) > f(\mathbf{G}^{best}(k))$ then
        • set $\mathbf{G}^{best}(k+1) = \boldsymbol{\Lambda}_j(k+1))$
      end if
    end if
  end for
end for
• obtain the final result, $\boldsymbol{\Lambda}^*$, by selecting from $\mathbf{G}^{best}(K)$.
```

It is worth mentioning that the direction of each particle move is determined by comparing the current energy with the present local/global optima. Thus, while the magnitude of the move is chosen randomly, the direction is not. By doing so, PSO discourages the solution from becoming trapped in local optima by moving each particle in a way that considers both the local and global best solutions from among all current particles. All particle hypotheses are clipped to be within the range of the CVIM latent space, and the maximum velocity is restricted within ±10% of the range of the latent space [148]. We summarize the PSO algorithm in Algorithm 2.

### 3.5 Gradient-boosted Particle Swarm Optimization

The PSO algorithm is simple, straightforward and robust, but it suffers high computational load due to CVIM-based shape interpolation, as well as the large number of iterations that are typically needed to obtain convergence. In some applications, the gradient is incorporated in sampling optimization to achieve a higher convergence rate [55]. In this section, we take advantage of the parametric nature of CVIM and incorporate a gradient-ascent step in the PSO to obtain a gradient-boosted PSO (GB-PSO) that overcomes these limitations by balancing between exploration and convergence with a deterministic and fast local search. Thus, GB-PSO is expected to be both more efficient and effective than the basic PSO in Algorithm 2.

A classical gradient ascent method starts from an initial hypothesis in the search space, *i.e.*, the parameter space of CVIM denoted by $\mathbf{\Lambda}$; then, by computing the local gradient direction, small steps are made toward the maximum iteratively. Due to the smooth and continuous nature of CVIM, which generates the shape embedding function $\mathbf{\Phi}$, $f = P(\mathbf{\Lambda}|\Omega)$ can be differentiated with respect to $\mathbf{\Lambda}$. Beginning from some initial guesses $\mathbf{\Lambda}_0$, we will then update our guess iteratively along the gradient direction:

$$\mathbf{\Lambda}^{(k+1)} = \mathbf{\Lambda}^{(k)} - r \cdot (-\nabla f|_{\mathbf{\Lambda}=\mathbf{\Lambda}^{(k)}}), = \mathbf{\Lambda}^{(k)} + r \cdot \nabla f|_{\mathbf{\Lambda}=\mathbf{\Lambda}^{(k)}} \tag{3.27}$$

where $r$ is the learning rate that determines the step size and $\nabla f|_{\mathbf{\Lambda}=\mathbf{\Lambda}_k}$ is the gradient of $f$ evaluated at the old guess. To compute $\nabla f|_{\mathbf{\Lambda}=\mathbf{\Lambda}_k}$, we take the derivative of $f$ with respect to $\mathbf{\Lambda}$ by the chain rule as:

$$\frac{\partial f}{\partial \mathbf{\Lambda}} = \left(\left(\frac{\partial f}{\partial \mathbf{\Phi}}\right)^T \cdot \frac{\partial \mathbf{\Phi}}{\partial \mathbf{\Lambda}}\right)^T = \left(\frac{\partial \mathbf{\Phi}}{\partial \mathbf{\Lambda}}\right)^T \frac{\partial f}{\partial \mathbf{\Phi}} \tag{3.28}$$

Similar to [10], the first term in Equation (3.28) can be written as:

$$\frac{\partial f}{\partial \mathbf{\Phi}} = \frac{\delta_\epsilon(\mathbf{\Phi})(P_f - P_b)}{H_\epsilon(\mathbf{\Phi})P_f + (1 - H_\epsilon(\mathbf{\Phi}))P_b} \tag{3.29}$$

where $\delta_\epsilon(\cdot)$ is the derivative of the Heaviside step function and $P_f$ and $P_b$ are defined in Equations (3.16) and (3.17). Since the latent variable $\mathbf{\Lambda} = [\mathbf{\Theta}^T, \alpha]^T$, so the second term in Equation (3.28) may be written as:

$$\frac{\partial \mathbf{\Phi}}{\partial \mathbf{\Lambda}} = \begin{bmatrix} \mathsf{IDCT}(\frac{\partial \mathcal{S}_{\text{DCT}}}{\partial \mathbf{\Theta}})^T \\ \mathsf{IDCT}(\frac{\partial \mathcal{S}_{\text{DCT}}}{\partial \alpha})^T \end{bmatrix}^T \tag{3.30}$$

The CVIM-based DCT generation of $\mathcal{S}_{\text{DCT}}$ is defined in Equation (3.2). From the properties of the tensor multiplication [149], we can rewrite Equation (3.2) as:

$$\mathcal{S}_{\text{DCT}}(\alpha, \mathbf{\Theta}) = \mathbf{A} \times_3 \mathbf{i}(\alpha) \times_2 \mathbf{\Psi}(\mathbf{\Theta}) = \mathbf{A} \times_2 \mathbf{\Psi}(\mathbf{\Theta}) \times_3 \mathbf{i}(\alpha) \tag{3.31}$$

where both $\mathbf{i}(\alpha)$ and $\Psi(\mathbf{\Theta})$ are row vectors. The steepest ascent optimization may then be performed based on the gradients along the view and identity manifolds.

### (I) Gradient along the view manifold

Let $\mathbf{B} = \mathbf{A} \times_3 \mathbf{i}(\alpha)$. From the tensor multiplication and flattening properties [149], it then follows that:

$$\mathcal{S}_{\text{DCT}}(\alpha, \mathbf{\Theta}) = \mathbf{B} \times_2 \Psi(\mathbf{\Theta}) = \mathbf{B}_{(2)}^T \cdot \Psi^T(\mathbf{\Theta}) \tag{3.32}$$

where $\mathbf{B}_{(2)}$ is the mode-two flattened matrix of $\mathbf{B}$. Hence,

$$\frac{\partial \mathcal{S}_{\text{DCT}}}{\partial \mathbf{\Theta}} = \mathbf{B}_{(2)}^T \cdot \frac{\partial \Psi^T(\mathbf{\Theta})}{\partial \mathbf{\Theta}} \tag{3.33}$$

It then follows from Equation (3.4) that:

$$\frac{\partial \Psi^T(\mathbf{\Theta})}{\partial \mathbf{\Theta}} = -2c[\kappa(\|\mathbf{\Theta} - \mathbf{S}_1\|)(\mathbf{\Theta} - \mathbf{S}_1), \cdots, \kappa(\|\mathbf{\Theta} - \mathbf{S}_{N_c}\|)(\mathbf{\Theta} - \mathbf{S}_{N_c})]^T \tag{3.34}$$

where $\kappa(\|\mathbf{\Theta} - \mathbf{S}_i\|)$ is defined in Equation (3.4). For the first term in Equation (3.30), we then have:

$$\frac{\partial \mathcal{S}_{\text{DCT}}}{\partial \mathbf{\Theta}} = -2c \cdot \mathbf{B}_{(2)}^T \cdot [\kappa(\|\mathbf{\Theta} - \mathbf{S}_1\|)(\mathbf{\Theta} - \mathbf{S}_1), \cdots, \kappa(\|\mathbf{\Theta} - \mathbf{S}_{N_c}\|)(\mathbf{\Theta} - \mathbf{S}_{N_c})]^T \tag{3.35}$$

**(II) Gradient along the identity manifold**

Let $\mathbf{C} = \mathbf{A} \times_2 \Psi(\mathbf{\Theta})$. From Equation (3.31), we have then that:

$$\mathcal{S}_{\text{DCT}}(\alpha, \mathbf{\Theta}) = \mathbf{C} \times_3 \mathbf{i}(\alpha) = \mathbf{C}_{(3)}^T \cdot \mathbf{i}^T(\alpha) \tag{3.36}$$

so:

$$\frac{\partial \mathcal{S}_{\text{DCT}}}{\partial \alpha} = \mathbf{C}_{(3)}^T \cdot \frac{\partial \mathbf{i}^T(\alpha)}{\partial \alpha} \tag{3.37}$$

Since $\mathbf{i}(\alpha)$ is the piecewise polynomial interpolation function, which is differentiable between any two given data points, it follows from Equation (3.3) that:

$$\frac{\partial \mathbf{i}^T(\alpha)}{\partial \alpha} = 3\mathbf{a}_m^T(\alpha - \alpha_m)^2 + 2\mathbf{b}_m^T(\alpha - \alpha_m) + \mathbf{c}_m^T, \quad \alpha \in [\alpha_m, \alpha_{m+1}) \tag{3.38}$$

Thus, we obtain finally:

$$\frac{\partial \mathcal{S}_{\text{DCT}}}{\partial \alpha} = \mathbf{C}_{(3)}^T \cdot [3\mathbf{a}_m^T(\alpha - \alpha_m)^2 + 2\mathbf{b}_m^T(\alpha - \alpha_m) + \mathbf{c}_m^T], \quad \alpha \in [\alpha_m, \alpha_{m+1}) \tag{3.39}$$

which together with Equation (3.35) provides the explicit formulation for both terms in Equation (3.30).

**(III) Gradient in the latent space**

From Equations (3.29) and (3.30), $\nabla f|_{\mathbf{\Lambda}=\mathbf{\Lambda}_k}$ may be rewritten as:

$$\nabla f|_{\mathbf{\Lambda}=\mathbf{\Lambda}_k} = \begin{bmatrix} \mathsf{IDCT}(\frac{\partial \mathcal{S}_{\text{DCT}}}{\partial \mathbf{\Theta}_k})^T \\ \\ \\ \mathsf{IDCT}(\frac{\partial \mathcal{S}_{\text{DCT}}}{\partial \alpha_k})^T \end{bmatrix} \frac{\delta_\epsilon(\mathbf{\Phi}_k)(P_f - P_b)}{H_\epsilon(\mathbf{\Phi}_k)P_f + (1 - H_\epsilon(\mathbf{\Phi}_k))P_b} \tag{3.40}$$

where $\alpha \in [\alpha_m, \alpha_{m+1})$, $\mathbf{\Phi}_k = \mathbf{\Phi}(\mathbf{\Lambda}_k)$ and $P(\mathbf{x}|\mathbf{\Lambda}_k, \mathbf{y})$ are defined in Equations (3.15) and (3.5), while $\frac{\partial \mathcal{S}_{\text{DCT}}}{\partial \mathbf{\Theta}_k}$ and $\frac{\partial \mathcal{S}_{\text{DCT}}}{\partial \alpha_k}$ are defined in Equations (3.35) and (3.39).

**Algorithm 3** The gradient-boosted (GB)-PSO method.

```
Initialization
• refer to Algorithm 2 Lines 1 ~ 4
GB-PSO step
for each iteration step ($k = 0 : K - 1$) do
    for each particle ($j = 1 : ps$) do
        • refer to Algorithm 2 Lines 8 ~ 15
    end for
    Gradient Ascent Local Search
    • set $\mathbf{\Lambda}^0 = \mathbf{G}^{best}(k+1)$ for gradient ascent local search;
    for each gradient ascent step ($l = 1 : pl$) do
        • calculate $\nabla f|_{\mathbf{\Lambda}=\mathbf{\Lambda}^{l-1}}$
        if the gradient is not significant then
            break
        end if
        • draw a sample for step size $r$
        • update $\mathbf{\Lambda}^l = \mathbf{\Lambda}^{l-1} + r \cdot \nabla f|_{\mathbf{\Lambda}=\mathbf{\Lambda}^{l-1}}$ according to Equation (3.40)
    end for
    • set $\mathbf{G}^{best}(k+1) = \mathbf{\Lambda}^l$;
end for
• obtain the final result, $\mathbf{\Lambda}^*$, by selecting from $\mathbf{G}^{best}(K)$.
```

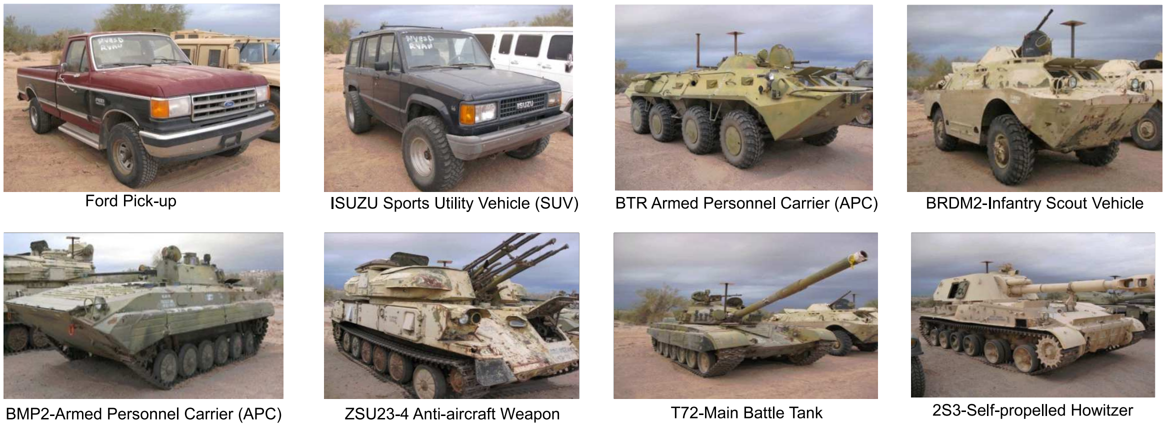
As suggested in [150], a uniformly-distributed random step size ($r$) could be used for the steepest ascent method, which turned out to be effective in practice. In practice, $r$ is uniformly distributed between $[\frac{\pi}{90}, \frac{\pi}{15}]$. In the GB-PSO method, the standard PSO is involved as the first step, then the global optimum ($\mathbf{G}^{best}(k+1)$) is updated by the gradient ascent method, which helps the next round PSO converge fast by improving velocity estimation. Thus, in the GB-PSO method, the total number of iterations required could be dramatically reduced compared with PSO. The computational load of the additional steps in GB-PSO is negligible due to two reasons: (1) the analytical nature of the energy function makes the gradient computation very efficient for the present global solution $\mathbf{G}^{best}(k+1)$ that is to be shared by all particles in the next iteration; (2) the update along the gradient direction is done analytically according to Equation (3.40), and there is a maximum number of gradient ascent iterations (*i.e.*, $pl = 20$ in this work) and a check of the current gradient value to determine if additional moves are necessary. In our experiment, we found that the actual number of the steps along the gradient direction is often much less than $pl$ (around 10), which confirms that the solution of gradient-based search is in the proximity of a local optimum. We summarize the GB-PSO method in Algorithm 3.

## 3.6 Experimental Results

In this work, our interest is to develop a general model-driven ATR algorithm where no IR data are used for training and no prior feature extraction is needed from the IR data, unlike most traditional methods that heavily rely on the quality of the training data, as well as feature extraction. We have conducted two comparative studies to evaluate the performance of the proposed algorithms. First, we have involved five comparable algorithms, including the proposed PSO and GB-PSO algorithms, all of which apply CVIM for shape modeling. Second, we also compared our algorithms with several recent ATR algorithms, including two SRC-based approaches [11, 103] and our previously-proposed ATR algorithm, which involves a joint view-identity manifold (JVIM) for target tracking and recognition [53]. The purpose of the first study is to validate the advantages of "implicit shape matching" over "explicit shape matching", as well as the efficiency of GB-PSO over PSO. That of the second study is to demonstrate the effectiveness of our new ATR algorithms compared with the recent ones in a similar experimental setting. It was reported in [4] that SRC-based methods can achieve state-of-the-art performance. In the following, we will first discuss the experimental setup shared by two comparative studies along with the metrics used for performance evaluation. Then, we present the two comparative studies one-by-one in detail.

### 3.6.1 SENSIAC Data and Experimental Setup

Similar to [9], we selected six 3D CAD models for each of the six target classes for CVIM training (36 models in total; Figure 3.2): APCs (armored personnel carriers), tanks, pick-ups, sedans, vans and SUVs. We considered elevation angles in $0° \sim 40°$ and azimuth angles in $0° \sim 360°$, with 10° and 12° intervals along the elevation and azimuth angles, respectively, on the view manifold, giving 150 multi-view shapes for each target. We also adopted a DCT-based shape descriptor [52], which facilitates CVIM learning and shape inference.


Ford Pick-up

ISUZU Sports Utility Vehicle (SUV)

BTR Armed Personnel Carrier (APC)

BRDM2-Infantry Scout Vehicle

BMP2-Armed Personnel Carrier (APC)

ZSU23-4 Anti-aircraft Weapon

T72-Main Battle Tank

2S3-Self-propelled Howitzer

Figure 3.11: All eight targets we used for algorithm evaluation were from the SENSIAC database.

All experiments were performed against the SENSIAC database [3], which provides a large collection of mid-wave IR and visible data depicting seven military targets and two civilian vehicles. We used 24 mid-wave (23 night-time and 1 day-time) IR sequences captured from 8 targets (Figure 3.11) at 1 km, 2 km and 3 km. In each sequence, there is a civilian or military vehicle traversing a closed-circular path with a diameter of 100 m.

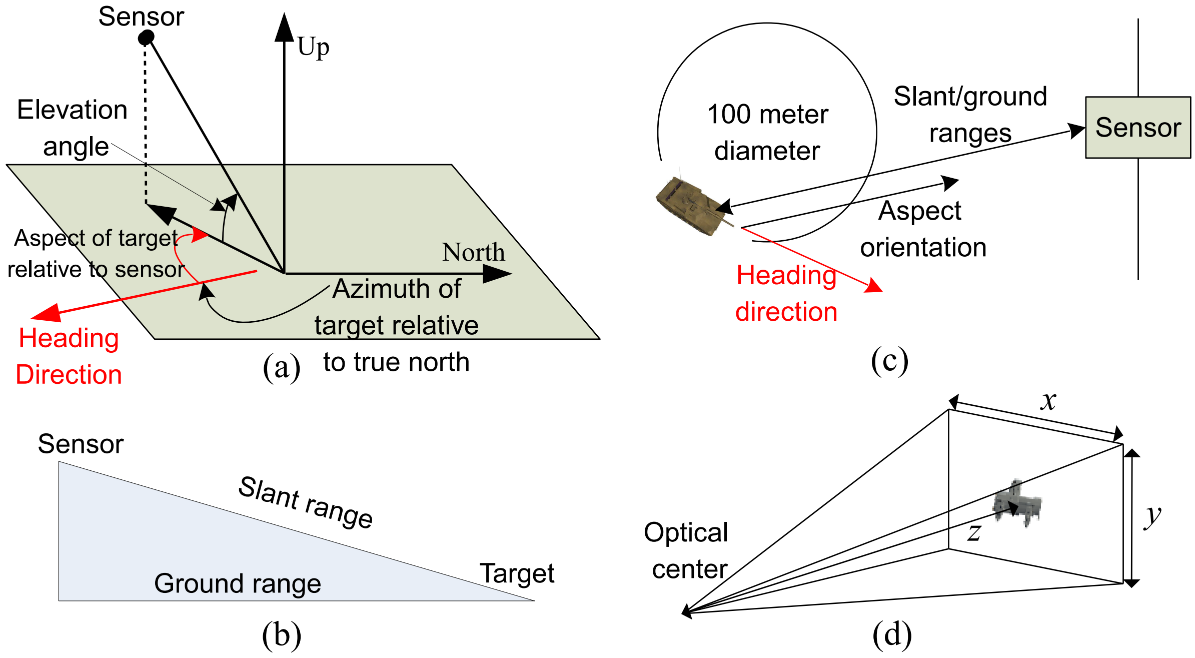


Figure 3.12: Definition of 3D coordinate system and the spatial geometry of the sensor, and also the target in the SENSIAC database.(a) The aspect of the target relative to the sensor, and the azimuth angle of target relative to the true north, and the elevation angle of the sensor; (b) side view of the ground and slant distances between the target and sensor; (c) top-down view of the aspect direction and the heading orientation; (d) a sensor-centroid 3D coordinate system.

We selected 100 frames from each sequence by down-sampling each sequence that has 1800 frames originally, where the aspect angle ranges from 0° to 360° with around a $5^\circ - 10^\circ$ interval; so in total, there are 2400 frames used for evaluation. The SENSIAC database also provides a rich amount of metadata (Figure 3.12), which can be used for performance evaluation, such as the aspect angle of the target, the field of view and the $2D$ bounding box of the target in each frame. Since we mainly focus on the recognition rather than detection, we also generated our target chips with the help of target 2D locations from this metadata (averaging around $50 \times 30 = 1500$ pixels at 1 km, $25 \times 14 = 350$ pixels at 2 km and $15 \times 10 = 150$ pixels at 3 km) in our experiments.

### 3.6.2 Comparative Study on Shape Modeling-based Approaches

This study compares five CVIM-based algorithms, which involve different segmentation and optimization techniques. Specifically, Method I uses background subtraction [151] to pre-segment a target-of-interest. This method is only suitable for a stationary sensor platform. Method II applies level set segmentation without a shape prior [10]. Both Method I and Method II need explicit shape matching, which involves Markov Chain Monte Carlo (MCMC)-based CVIM inference after segmentation to accomplish ATR [57]. Method III applies a multi-threaded MCMC-based inference technique to jointly optimize over CVIM in a level set by involving implicit shape matching without target pre-segmentation. It was shown in [57] that Method III significantly outperforms the first two, but it suffers from high computational complexity due to the MCMC-based shape inference. PSO and GB-PSO are referred to as Methods IV and V, respectively. The computational time for each ATR chip ($50 \times 30$ pixels) for five algorithms is around 10, 14, 22, 15 and 6 s, respectively, using an un-optimized MATLAB code on a PC with a Quad-core CPU (2.5 GHZ).

We evaluate these five algorithms with respect to: (1) the accuracy of pose estimation (*i.e.*, the aspect angle); (2) the 2D pixel location errors between the segmented shape and the ground truth bounding box; (3) the recognition accuracy in terms of six major target classes; and (4) the sensor-target distance (*i.e.*, range, computed by scaling factors) errors in meters. To examine the robustness of our algorithms, we analyze (5) the recognition accuracy *versus* three related factors, *i.e.*, the contrast of image chips [152], the foreground/background $\chi^2$ histogram distance [153] based on the segmentation results and the aspect angle. The chip contrast and $\chi^2$ histogram distance indicate the IR image quality and the target visibility, respectively. Similar to [154–157], we also evaluate the overlap ratio between the estimated bounding box (derived from the segmentation result) and the ground truth bounding box (available from ground-truth data) (6), which is a simple, yet effective and widely-accepted way to quantify the segmentation performance. Furthermore, we manually created the ground truth segmentation masks from five randomly-selected frames per IR sequence, so that we can compute the overlap ratio between the segmentation results with the ground-truth masks (7). Moreover we will show the capability of the proposed algorithm (GB-PSO) for sub-class recognition, *i.e.*, the specific target type within a class, even if the exact target type is not in the training data.

*(i) Pose Estimation Results*

Table 3.1 reports aspect angle error (pose estimation) results for all five tested methods along with the 2D pixel error and 3D range error in the predefined 3D camera coordinate system (given in the metadata).

Table 3.1: Pose and location estimation errors for all five tested methods averaged over SENSIAC mid-wave IR sequences for each target to sensor distances depicting eight difference targets (Method I/Method II/Method III/Method IV/Method V).

| | **2D Pixel Error (pixels)** | **Aspect Angle Error (°)** | **Range Error (m)** |
|---|---|---|---|
| 1 km | 2.8/3.1/1.9/2.1/**1.9** | 17.2/17.9/15.1/**14.8**/15.1 | 25.1/27.8/24.2/**24.1**/24.5 |
| 2 km | 2.9/3.4/2.3/2.4/**2.1** | 21.2/25.2/18.7/18.2/**17.1** | 39.1/38.2/33.8/32.8/**32.6** |
| 3 km | 2.5/3.8/2.2/**1.8**/2.0 | 26.1/27.5/21.7/21.9/**20.5** | 43.5/48.3/40.2/**40.1**/41.1 |

We can see clearly that both Methods IV and V can achieve moderate, significant and slight improvements over Method I (background subtraction for segmentation), Method II (level set segmentation without shape prior) and Method III (MCMC-based CVIM inference), respectively. Although Methods IV and V do not provide a significant improvement in pose and location estimation performance compared to Method III, they provide similar performance at a greatly reduced computational

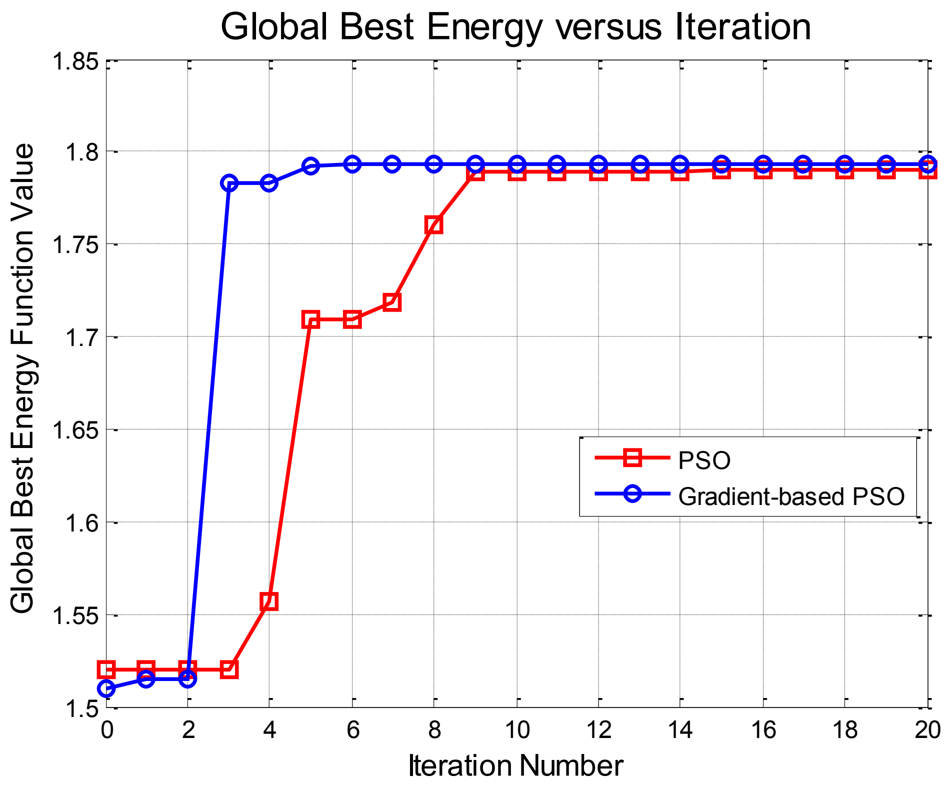


Figure 3.13: Convergence plots for PSO and GB-PSO in one example.

complexity. Numerical results from PSO and GB-PSO are comparable to each other. However, Figure 3.13 shows that GB-PSO converges nearly three-times faster than the PSO, demonstrating the value of gradient boosting in the CVIM latent space.

*(ii) Target Recognition Results*

The recognition results are computed based on the percentage of frames where the target class is correctly classified. As shown in Table 3.2, both PSO (Method IV) and GB-PSO (Method V) generally achieve modest performance gains over Method I–III, while GB-PSO does so with a significantly reduced computational complexity compared to all four of the other methods. Furthermore, Figure 3.14 shows some sub-class recognition results for eight 1-km IR images. The sub-class recognition can be

Table 3.2: Overall recognition and segmentation results of five methods (I /II /III /IV /V).

| | **Average Recognition** | **Bounding Box** |
|---|---|---|
| | **Accuracy (%)** | **Overlap (%)** |
| 1 km | 81/78/85/**86**/85 | 85.2/82.9/88.1/88.6/**88.9** |
| 2 km | 71/64/73/75/**76** | 75.6/74.1/79.5/**79.8**/79.2 |
| 3 km | 69/62/70/72/**73** | 67.7/65.5/70.1/70.9/**71.6** |
| | **Segmentation Mask** | **Fore/Background** |
| | **Overlap (%)** | $\chi^2$ **Histogram Distance** |
| 1 km | 79.3/83.2/83.5/**83.8**/83.6 | 0.30/0.32/0.34/0.34/**0.35** |
| 2 km | 72.3/73.9/79.8/79.6/**80.1** | 0.26/0.25/0.29/0.28/**0.31** |
| 3 km | 63.8/67.2/68.8/**69.3**/69.1 | 0.20/0.23/0.25/**0.26**/0.25 |

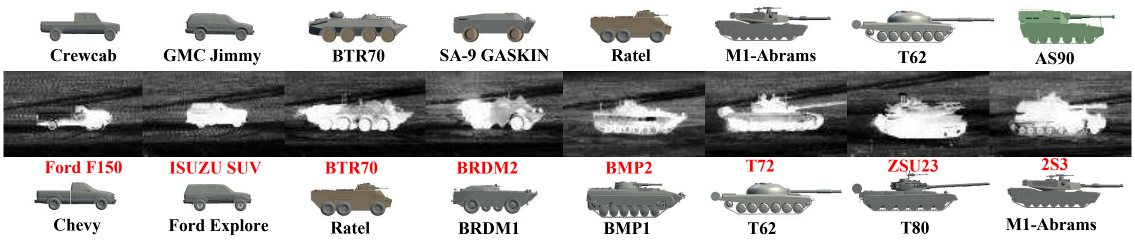


Figure 3.14: Some sub-class recognition results under 1000 m. The first row and third row show the closest training vehicles along the identity manifold in the CVIM, and the middle row presents the original IR chips.

achieved via CVIM by finding the two closest training target types along the identity manifold. Since the training data only have the BTR70 model, we find that we can recognize the BTR70 at the sub-class level most of the time. Interestingly, we can see that T72, BMP2 and 2S3 are also recognized as T62, BMP1 and AS90, respectively, which are the closest sub-class target types available in our training data.

We also summarize the recognition performance with respect to image contrast and fore/background histogram distance in Figure 3.15a and b. Our algorithm tends to perform worse when encountering frames with a lower contrast ratio and a smaller fore/background histogram distance. Furthermore, we demonstrate the effect of view variations on the recognition performances we well, as shown shown in Figure 3.15c and d. Most failure cases (red dots) occur at around 0° (or 360°) and 180° (front/rear views). Since we only use the shape information here, a more advanced target appearance representation that involves intensity and other features could make ATR performance more robust to view variation.

To make further demonstrations, we summarize recognition results from GB-PSO *vs.* the chip contrast, foreground/background histogram distance and aspect angle in Figure 3.16. It is shown in Figure 3.16a that our algorithm performs well for most chips with reasonable contrast and tends to deteriorate for chips with a very low contrast, which is usually associated with poor image quality (e.g., day-time IR imagery). As illustrated in Figure 3.16b, the foreground/background $\chi^2$ histogram distance is strongly related to the recognition accuracy. This is because the $\chi^2$ distance is related to the target visibility and can also quantify the segmentation quality. When segmentation results are good with large background/foreground separation (large $\chi^2$ distance values), the recognition accuracies are usually high, which also imply good target segmentations. Furthermore, the aspect angle is a key factor that affects the recognition performance. As shown in Figure 3.16c, the highest accuracy occurs around the side views (90° and 270°) when the targets are most recognizable. Most

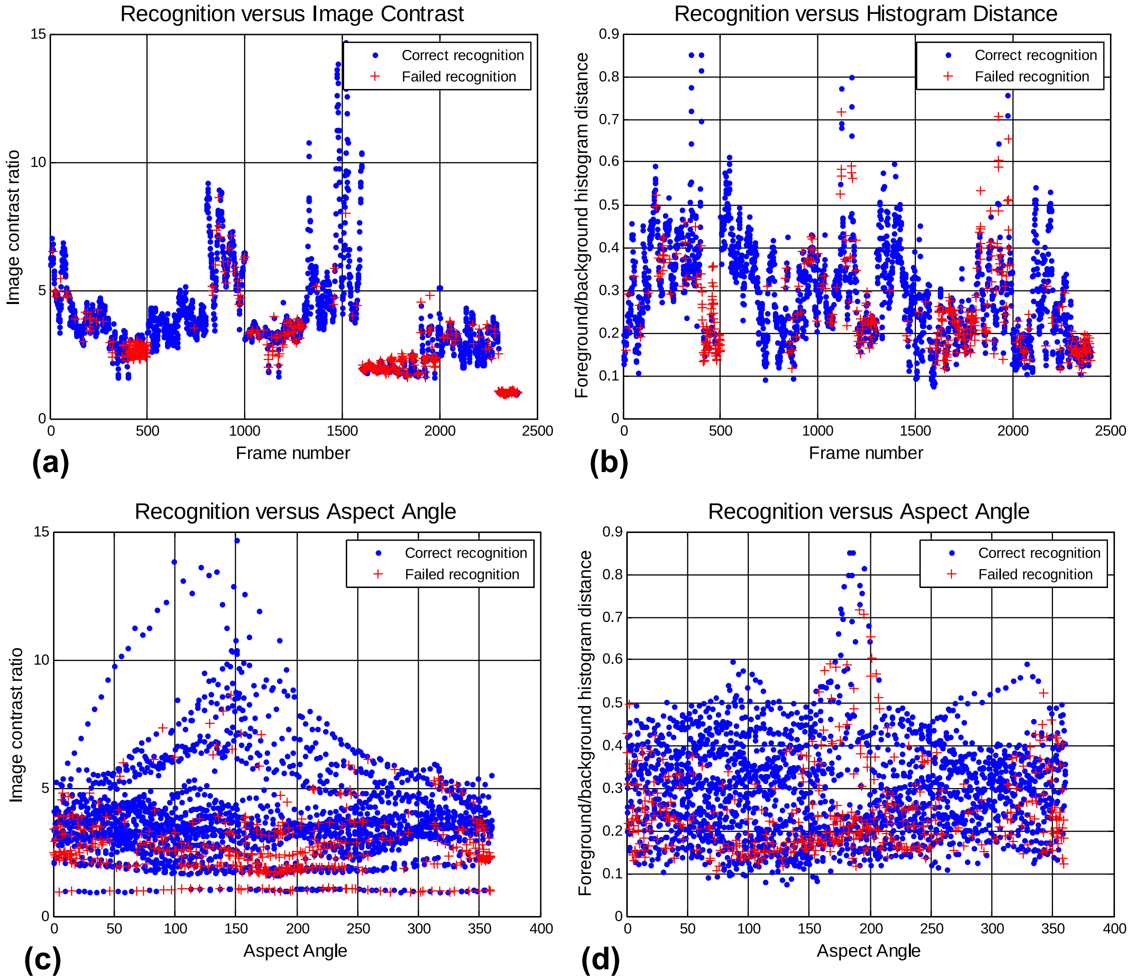
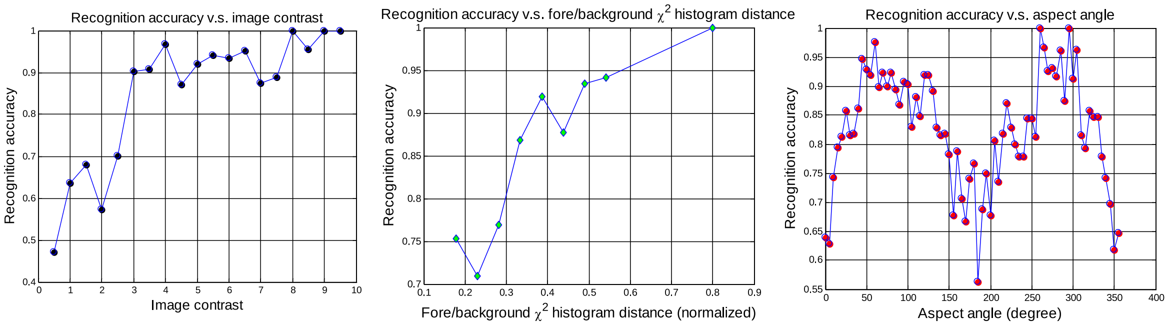


Figure 3.15: The experimental analysis of GB-PSO. (a) and (b): The image contrast ratio and fore/background histogram distance versus overall 2400 frames. (c) and (d) The image contrast ratio and fore/background histogram distance versus the aspect angle for overall 2400 frames. Blue dots are correctly recognized frames while red dots are mis-recognized frames. (This figure is better viewed in color.)



(a) (b) (c)

Figure 3.16: Robustness analysis of GB-PSO over all 2400 chips. (**a**) Recognition accuracy *versus* the chip contrast; (**b**) recognition accuracy *versus* fore/background histogram distances; (**c**) recognition accuracy *versus* the aspect angles.

failed cases are around 0° (or 360°) (frontal views) and 180° (rear views), when it is hard to differentiate different targets due to the high shape ambiguity. Since we only use the shape information here, a more advanced and informative target appearance representation that involves intensity and other features could make ATR performance more robust to aspect angles.

### *(iii) Target Segmentation Results*

Table 3.2 shows the target segmentation results in terms of the bounding box overlap, the segmentation mask overlap and the foreground/background $\chi^2$ histogram distance. Both PSO and GB-PSO outperform Methods I and II, while performing comparably to Method III at a lower computational complexity. Figure 3.17 shows some snapshots of the original IR imagery of eight targets under the 1-km range, along with the manually-cropped segmentation masks, the results of the background subtraction segmentation, the level set segmentation without a shape prior and the PSO method, respectively. It may be seen that the CVIM shape prior drives the segmentation to a semantically more meaningful shape compared to Methods I and II, where a shape prior is not involved.

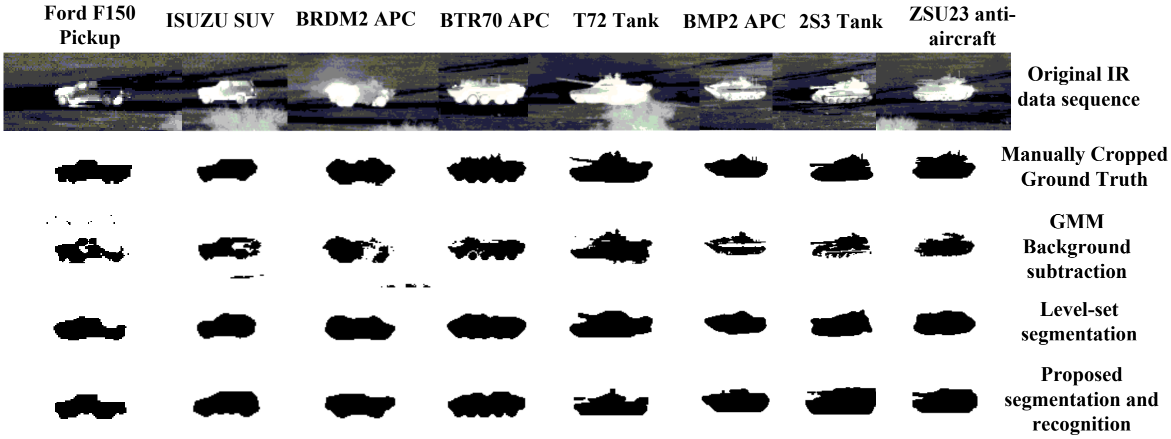

Ford F150 Pickup
ISUZU SUV
BRDM2 APC
BTR70 APC
T72 Tank
BMP2 APC
2S3 Tank
ZSU23 anti-aircraft
Original IR data sequence
Manually Cropped Ground Truth
GMM Background subtraction
Level-set segmentation
Proposed segmentation and recognition

Figure 3.17: Snapshot of the segmentation results. From the first row to the last: original IR frame, manually-cropped segmentation masks, results of background subtraction segmentation, level set segmentation without shape prior and the final segmentation and recognition results with CVIM shape prior (Method IV) interpolated from the CVIM.

Some snapshots of segmentation results along with pose estimation and recognition results of Method V (GB-PSO) are shown in Figure 3.18. It is found that GB-PSO is robust to background clutter and engine smoke, and the frontal/rear views may pose some challenge. For BRDM2 APC, we see a heat spot near the tail, which changes the target appearance significantly, but we can still recognize it as an APC. However, for targets of near front/rear views, although the segmentation results are still acceptable, the recognition results are often wrong. For example, the BMP2 APC was misrecognized as the M60 tank in the front view. By closely observing the IR appearances of BMP2 APC, we find that this particular APC does indeed look similar to a tank when viewed frontally. This can also be explained by our identity manifold topology learned by *class-constrained shortest-closed-path*, where the BMP2 stays closest to tanks along the identity manifold among all APCs, as shown in Figure 3.2.

A similar case happens to BTR70 APC. Moreover, the proposed algorithm as

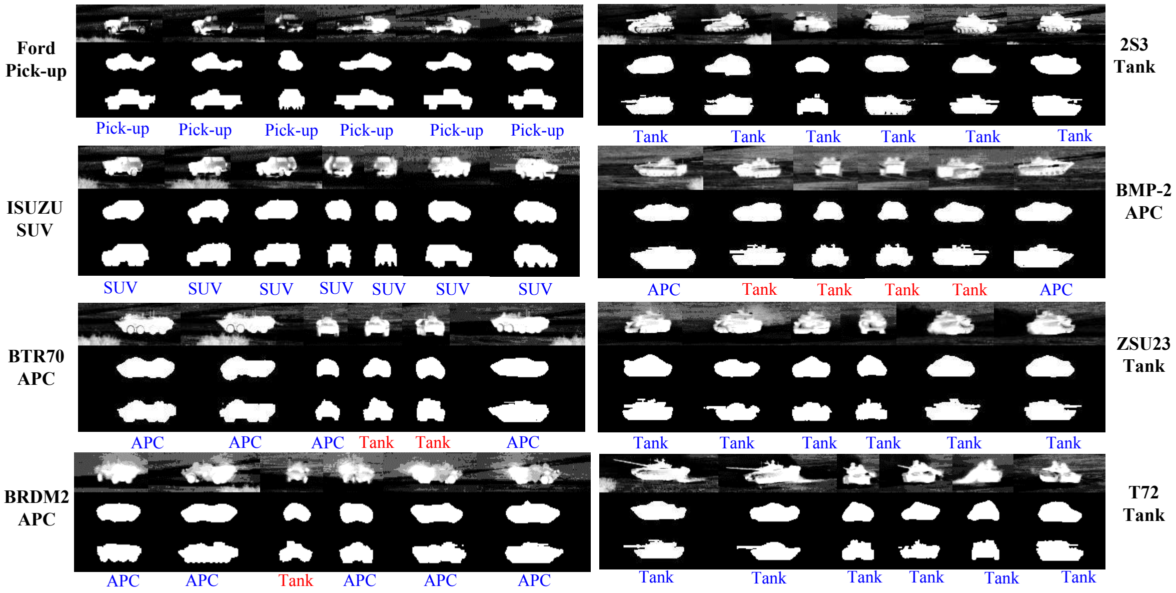


Figure 3.18: Some GB-PSO results for pose estimation and recognition results of eight targets under a 1-km target-sensor range from the SENSIAC database. For each target, from the first row to the third: original IR frame, level set segmentation-based initialization and the final pose estimation and recognition results from GB-PSO. Recognition results are denoted in blue if correct and red if wrong below each chip.

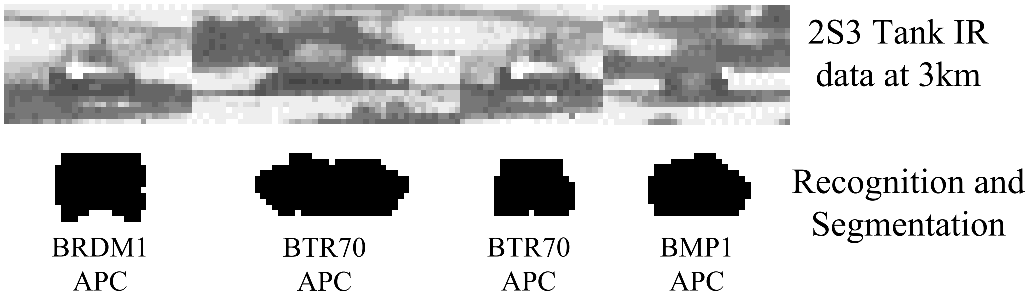


Figure 3.19: Some failed cases of the 2S3 tank in a day-time sequence at 3 km.

realized in Methods IV and V performs poorly against long-range day-time IR data (3 km), where the foreground/background contrast is low and the target is small. This is illustrated in Figure 3.19, where the 2S3 tank is misclassified as an APC in several frames. As we already mentioned, a more powerful appearance representation is needed to handle challenging cases of this type.

### 3.6.3 Comparative Study: Recent ATR Methods

This comparative study includes three recent ATR algorithms that are compared against PSO and GB-PSO. Specifically, we applied the gradient-based optimization technique discussed in [53] to apply JVIM (learned from the same set of training shapes as CVIM) for image-based ATR where level set segmentation without a shape prior is used for initialization. In addition, we have also implemented the SRC-based algorithm [4] and the multi-attribute Lasso with group constraint (MALGC) [11], which is an extended SRC approach by taking advantage of the attribute information (*i.e.*, angles) during sparse optimization. Both SRC algorithms require a dictionary that includes training shapes also used for CVIM learning. The input is the level set segmentation output from an IR chip. We also use the segmentation results from GB-PSO for SRC-based ATR (namely SRC-GB-PSO) to see the effect of good target segmentation. The computational time (in the same computational setting as before) for four implementations is around 4 (JVIM), 16 (SRC), 20 (MALGC) and 18 (SRC-GB-PSO) s, compared with 15 and 6 s for PSO and GB-PSO, respectively. We compare six ATR algorithms in Table 3.3, and we have the following observations and discussion according to Table 3.3.

- The JVIM model unifies the view and identity manifolds in one latent space for more accurate shape modeling than CVIM, which involves separate view and identity manifolds [158], and it is especially suitable for target tracking due to the unified and smooth shape manifold [53]. However, JVIM has a simi-

Table 3.3: The performance comparison with recent ATR methods in terms of the recognition accuracy and the aspect angle error (joint view-identity manifold (JVIM)/sparse representation-based classification (SRC)/multi-attribute Lasso with group constraint (MALGC)/ SRC-GB-PSO/ PSO/ GB-PSO).

| **Ranges** | **Average Recognition Accuracy (%)** | **Aspect Angle Error (°)** |
|---|---|---|
| 1 km | 82/83/83/85/**86**/85 | 16.9/15.6/15.4/15.3/**14.8**/15.1 |
| 2 km | 69/72/73/75/75/**76** | 22.3/20.1/19.9/17.8/18.2/**17.1** |
| 3 km | 65/69/70/72/72/**73** | 26.8/24.5/23.9/21.1/21.9/**20.5** |

lar multi-modal problem as CVIM that makes the gradient-based optimization often trapped in local minima, as reflected by relatively poor results in Table 3.3. It may not be efficient to apply sample-based approaches (e.g., PSO or MCMC) to JVIM optimization due to the fact that its joint manifold structure will require a large number of samples to ensure effective sampling. The reason that JVIM shows promising results in target tracking is because the dynamic modeling involved greatly facilitates sequential state estimation. In the case of image-based ATR, CVIM shows some advantages over JVIM due to its simpler structure.

- Both SRC and MALGC methods show reasonable performance by only using shapes segmented by level set for ATR. Especially for the range of 1 km when target segmentation is likely more accurate, all algorithms are very comparable. It is observed that MALGC is slightly better than SRC by utilizing the angle information in sparse optimization. More interestingly, we can see that better target segmentation results via GB-PSO moderately improve the SRC's performance (the fourth algorithm, GB-PSO-SRC). The computational complexity of SRC and MALGC is comparable with PSO and much higher than that of GB-PSO.

- It is shown that the proposed PSO and GB-PSO algorithms are comparable, both of which are better than the others in all cases. Although the improvement of GB-PSO over the other two SRC approaches is moderate, it does offer several advantages: (1) it is computationally efficient with only a 30%–40% computational load; (2) target segmentation is not required and can be considered as a byproduct of ATR; and (3) the proposed GB-PSO algorithm is a model-driven approach that does not require real-world training data, and it has potential to be combined with other data-driven approaches to further improve ATR performance.

## 3.7 Discussion

In this chapter we presented an a probabilistic level set framework with shape generative modeling for joint target recognition, segmentation and pose estimation for IR imagery. Due to the multi-mode nature of the energy function, we propose a multi-threaded Monte Carlo Markov Chain (MCMC)-based strategy, and we also adopted the particle swarm optimization (PSO) algorithm and a gradient-boosted PSO (GB-PSO) for faster optimization. Experimental results on the SENSIAC ATR dataset demonstrate the effectiveness of the proposed framework. Both PSO algorithms (especially GB-PSO) reduce the cost of shape matching during CVIM-based shape inference, and they are more efficient and effective than other traditional implementations which require intensive shape matching either explicitly (with pre-segmentation) or implicitly (without pre-segmentation). The proposed framework could be further extended to incorporate advanced appearance representations when target shape observations are not preferably detected. In the future we aim to build more comprehensive ATR systems assisted by global contextual information.

# CHAPTER IV

# SPARSE REPRESENTATION-BASED ATR

In the previous chapter we integrated a shape generative model (CVIM) into a probabilistic level set framework to implement joint target recognition, segmentation and pose estimation, and our algorithms perform well when target boundaries are recognizable with ease. But most failure cases occur when encountering highly clutters in the surrounding area or much complicated background conditions caused by dust or weather especially in the day time IR sequences, where even human eyes could spend hard time to recognize the target signature. In this chapter, we further explore some data-driven approaches on a more challenging dataset, where target boundaries are not preferably detected and insufficient shape features could be applied for shape modeling-based optimization approaches. Specifically, we will first introduce two recent sparse representation classification methods for recognition tasks, followed by our combination framework for an improved recognition performance.

## 4.1 Preliminary Work

As mentioned in Chapter II, sparse representation classification methods (SRC) has been successfully applied in human face recognition [103] and also ATR [4]. Some more advanced applications such as multi-attribute sparse representation based method (MALGC) was introduced in [11], which encourages group constraints on different attribute of data for the sparsity solution, such as lighting condition, facial expression, etc. In this and following sections we will briefly introduce some SRC methods and present some in-depth analysis on performances of sparse representation based recognition on the Comanche dataset.

### 4.1.1 Sparse Representation-based Classification

Suppose there are $L$ target classes, and each class has $n$ training images, while each image is $w \times h$, and we vectorize them as $N$-dimensional vectors, where $N = w \times h$. Let $\mathbf{A}_k = [\mathbf{D}_{k1}, ..., \mathbf{D}_{kn}]$ be an $N \times n$ matrix of training images from the $kth$ class as the library representing the $kth$ class, then an library that consists of all sub-libraries from all classes can be defined as

$$\begin{aligned} \mathbf{A} &= [\mathbf{A}_1, ..., \mathbf{A}_L] \in R^{N \times (n \cdot L)} \\ &= [\mathbf{D}_{11}, ..., \mathbf{D}_{1n} | \mathbf{D}_{21}, ..., \mathbf{D}_{2n} | ...... | \mathbf{D}_{L1}, ..., \mathbf{D}_{Ln}], \end{aligned} \tag{4.1}$$

Then, we can consider an observation vector $\mathbf{Y} \in R^N$ as a linear combination of all training data as

$$\mathbf{Y} = \sum_{i=1}^{L} \sum_{j=1}^{N} \alpha_{ij} \mathbf{D}_{ij}, \tag{4.2}$$

where $\alpha_{ij} \in R$. The equation above can be written in matrix form as

$$\mathbf{Y} = \mathbf{A}\boldsymbol{\alpha}, \tag{4.3}$$

where $\boldsymbol{\alpha} = [\alpha_{11}, ..., \alpha_{1n} | \alpha_{21}, ..., \alpha_{2n} | ...... | \alpha_{L1}, ..., \alpha_{Ln}]^T$. The assumption behind the sparse representation based classification method is that, suppose we are given sufficient training data, any observation that belong to the same class will approximately stay in the linear span of the training library from the same class - coefficients in $\boldsymbol{\alpha}$ that are not associated to the same class will be close to zero, which makes the $\boldsymbol{\alpha}$ a sparse vector. Researches have shown that if $\boldsymbol{\alpha}$ is sparse enough and certain properties applies, the sparsest $\boldsymbol{\alpha}$ can be found by solving the following optimization problems,

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}'} \|\boldsymbol{\alpha}'\|_1, \text{ subject to } \mathbf{Y} = \mathbf{A}\boldsymbol{\alpha}', \tag{4.4}$$

where $\| \cdot \|_1$ is the norm 1 operation. After this optimization, the idea case is that values of the coefficients $\hat{\boldsymbol{\alpha}}$ are mainly associated to columns of the library $\mathbf{A}$ from one single class, so based on this fact we are able to do a classification by comparing the reconstruction errors from different parts of the estimated $\hat{\boldsymbol{\alpha}}$ that merely belong to each class, and the minimum reconstruction errors is used to recognize the class of the observation. A function $\chi_k$ was introduced that can be used to select values of the coefficients associated with class $k$

$$\mathbf{r}_k(\mathbf{Y}) = \|\mathbf{Y} - \mathbf{A}\chi_k(\hat{\boldsymbol{\alpha}})\|_2, \tag{4.5}$$

then the class $d$ can be estimated as the one that produces smallest reconstruction error,

$$d = \arg\min_k \mathbf{r}_k(\mathbf{Y}), \tag{4.6}$$

where $d \in \{1, 2, ..., L\}$ is the classification result.

### 4.1.2 Multi-Attribute Sparse Representation

A simple and efficient multi-attribute Lasso with group constraint (MALGC), or multi-attribute sparse representation is proposed in [11] for improved classification accuracy on face recognition with jointly considering group constraints, reconstruction error and sparsity property. Consider that the training dataset may have different attributes, and each attribute can be represented as a binary matrix $\mathbf{A}_b \in R^{m_b \times (n \cdot L)}$, each of which has $m_b$ types, where $b$ is the index of attributes $b = 1, ..., B$. Take the Comanche data for example, it might contains three attributes: target types, view angles, illumination, and each attribute has may types, for example the target types contains ten different types of vehicles, and the view angles have 72 different degrees of view, while the illumination may contain day time and night time data. So each attribute is defined as

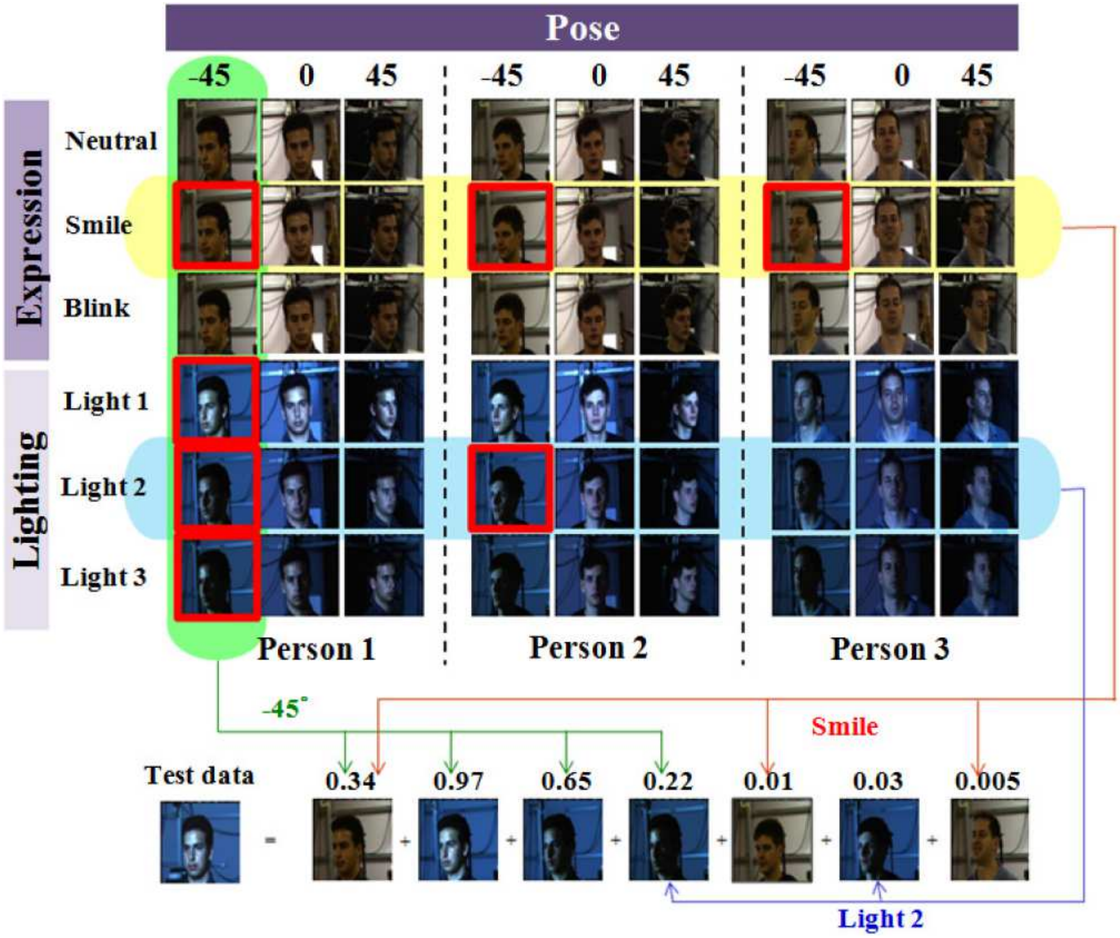


Figure 4.1: Multi-attribute sparse representation applied to human face recognition with face expression, pose and lighting attributes [11].

$$\mathbf{A}_b = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,p} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m_b,1} & a_{m_b,2} & \cdots & a_{m_b,p} \end{bmatrix}, \tag{4.7}$$

$$\text{where} \qquad a_{i,p} = \begin{cases} 1, & \text{if } \mathbf{D}_p \in \text{type } i \\ 0, & \text{otherwise} \end{cases} \tag{4.8}$$

and $i = 1, ..., m_b$ is the type index for attribute $b$, and each column in $A_b$ associate to the data $\mathbf{D}_p$, where $p = 1, ..., nL$. By adding all attribute as group constraints in

Equation (4.4), a new formulation of optimization problem is given,

$$\min_{\boldsymbol{\alpha}', \boldsymbol{\alpha}_{\boldsymbol{A_1}}', \ldots, \boldsymbol{\alpha}_{\boldsymbol{A_b}}'} \|\boldsymbol{\alpha'}\|_1 + \|\boldsymbol{\alpha'}_{\mathbf{A}_1}\|_1 + \|\boldsymbol{\alpha'}_{\mathbf{A}_2}\|_1 + \ldots + \|\boldsymbol{\alpha'}_{\mathbf{A}_B}\|_1,$$
$$\text{subject to } \mathbf{Y} = \mathbf{A}\boldsymbol{\alpha}', \boldsymbol{\alpha'}_{\mathbf{A}_1} = \mathbf{A_1}\boldsymbol{\alpha'}, \boldsymbol{\alpha'}_{\mathbf{A}_2} = \mathbf{A_2}\boldsymbol{\alpha'}, \ldots \boldsymbol{\alpha'}_{\mathbf{A}_B} = \mathbf{A_B}\boldsymbol{\alpha'}, \qquad (4.9)$$

obviously it is hard to solve such an optimization problem with many group constraints, but it can be reformulated to a typical $l_1$ optimization problem by simply reformulating it,

$$\min_{\tilde{\boldsymbol{\alpha}}} \|\tilde{\boldsymbol{\alpha}}\|_1, \text{ subject to } \tilde{\mathbf{Y}} = \tilde{\mathbf{A}}\tilde{\boldsymbol{\alpha}}, \qquad (4.10)$$

where

$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{0}_{N\times m_1} & \mathbf{0}_{N\times m_2} & \cdots & \mathbf{0}_{N\times m_B} \\ \mathbf{A}_1 & -\boldsymbol{I}_{m_1\times m_1} & \mathbf{0}_{m_1\times m_2} & \cdots & \mathbf{0}_{m_1\times m_B} \\ \mathbf{A}_2 & \mathbf{0}_{m_2\times m_1} & -\boldsymbol{I}_{m_2\times m_2} & \cdots & \mathbf{0}_{m_2\times m_B} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_B & \mathbf{0}_{m_B\times m_1} & \mathbf{0}_{m_B\times m_2} & \cdots & -\boldsymbol{I}_{m_B\times m_B} \end{bmatrix}, \qquad (4.11)$$

where $\boldsymbol{I}$ is the identity matrix, and

$$\tilde{\boldsymbol{\alpha}} = [\boldsymbol{\alpha}^T, \boldsymbol{\alpha}^T_{\mathbf{A_1}}, \boldsymbol{\alpha}^T_{\mathbf{A_2}}, \cdots, \boldsymbol{\alpha}^T_{\mathbf{A_B}}]^T,$$
$$\tilde{\mathbf{Y}} = [\mathbf{Y}^T, \mathbf{0}^T_{m_1\times 1}, \mathbf{0}^T_{m_2\times 1}, \cdots, \mathbf{0}^T_{m_B\times 1}]^T, \qquad (4.12)$$

therefore, Equation (4.10) can be easily solved by any $l_1$ optimization methods. The advantage here is that the multi-attribute sparse representation is able to offer more data concentration from various attributes as group constraints during optimization, and it can be reformulated as typical $l_1$ optimization problems as well.

Similar to most sparse representation classification problems, the final classification results are determined by the reconstruction error by the sparse coefficients vectors. The vector $\boldsymbol{\alpha}_{\mathbf{A}_b}$ has a dimension of $m_b$, which is the number of types in the $bth$ attribute, and each element in $\boldsymbol{\alpha}_{\mathbf{A}_b}$ would represent the sparse coefficients sum for each type in this attribute. a concept of basic group was introduced so that, for classification purposes, the target type is the basic attribute, while for view estimation purpose, the view angle is the basic attribute. In our work, the target ID is the basic attribute, as denoted by $\mathbf{A}_c$, and $\boldsymbol{\alpha}_{\mathbf{A}_c,q}$ is the *qth* element of vector $\boldsymbol{\alpha}_{\mathbf{A}_c}$ associated to the basic attribute matrix $\mathbf{A}_c$. Then the decision is based on the combination of reconstruction error of the observation and the group constraint,

$$d = \arg\max_{q} \big(e^{-\frac{\|\mathbf{Y}-\mathbf{A}\chi_q(\boldsymbol{\alpha})\|_2}{\sigma}} + \gamma \frac{|\boldsymbol{\alpha}_{\mathbf{A}_c,q}|}{\sum_q |\boldsymbol{\alpha}_{\mathbf{A}_c,q}|}), \tag{4.13}$$

where the $\chi_q(\boldsymbol{\alpha})$ select the coefficients in $\boldsymbol{\alpha}$ associated to the *qth* class. First term and second terms account for minimal reconstruction error and sparsest group constraint, respectively, and $\gamma$ is used here to control the balance between them.

## 4.2 The Comanche Dataset

As shown in Figure 4.2, we tested our Shape Manifold Aware level set method (Method V) on the Comanche dataset, which consists of ten targets, and there are 72 orientations for each target. We can see that only the first row the target signature is very clear, but from the 2nd to the last row, target signatures are much weaker comparing to the M60 tank. But our algorithm is robust enough as long as target boundaries are visually recognizable, such as the BMP APC, and the M60 tank. But in most cases we can achieve a fairly good recognition for main vehicle classes (Tanks, APCs, SUVs, Trucks). The recognition accuracy based on main classes is 78%, while subclass recognition drop down to 65%. But notice that there is no input data required for training.

Regarding to such challenging data where target signatures vary significantly, pure model-drive approaches might not be preferable in order to achieve higher recognition accuracy suppose enough training data are given. As reported in [4] a sparse representation method was applied to this Comanche data for target recognition, and

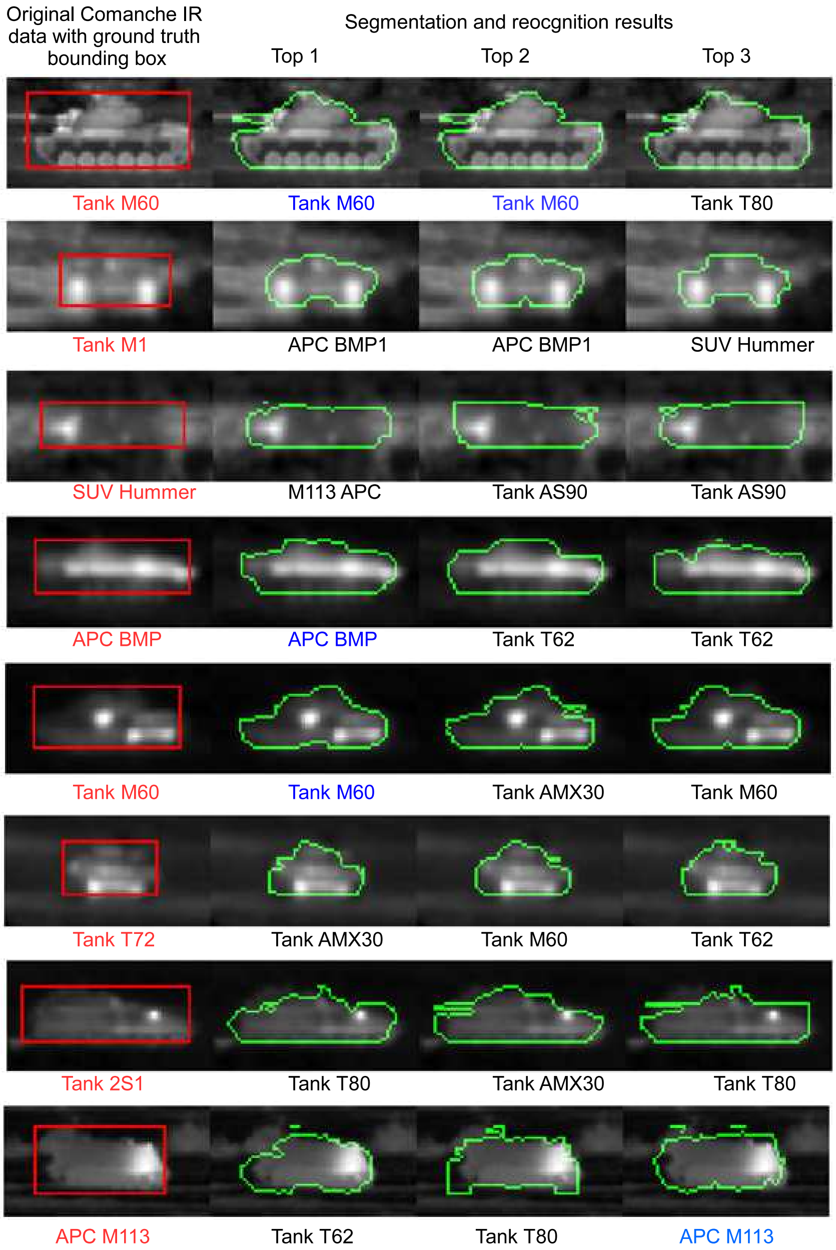


Figure 4.2: Segmentation and recognition results from Method-V. For each row, the first image is the image clip from the Comanche dataset with ground truth bounding box (red), and from the 2nd to the 4th we give the top three segmentation and recognition results respectively from left to right.

results outperform other traditional methods. In following sections we will explore our research on the sparse representation classification (SRC).

### 4.3 Switchable Sparse Representation Classification (SSRC)

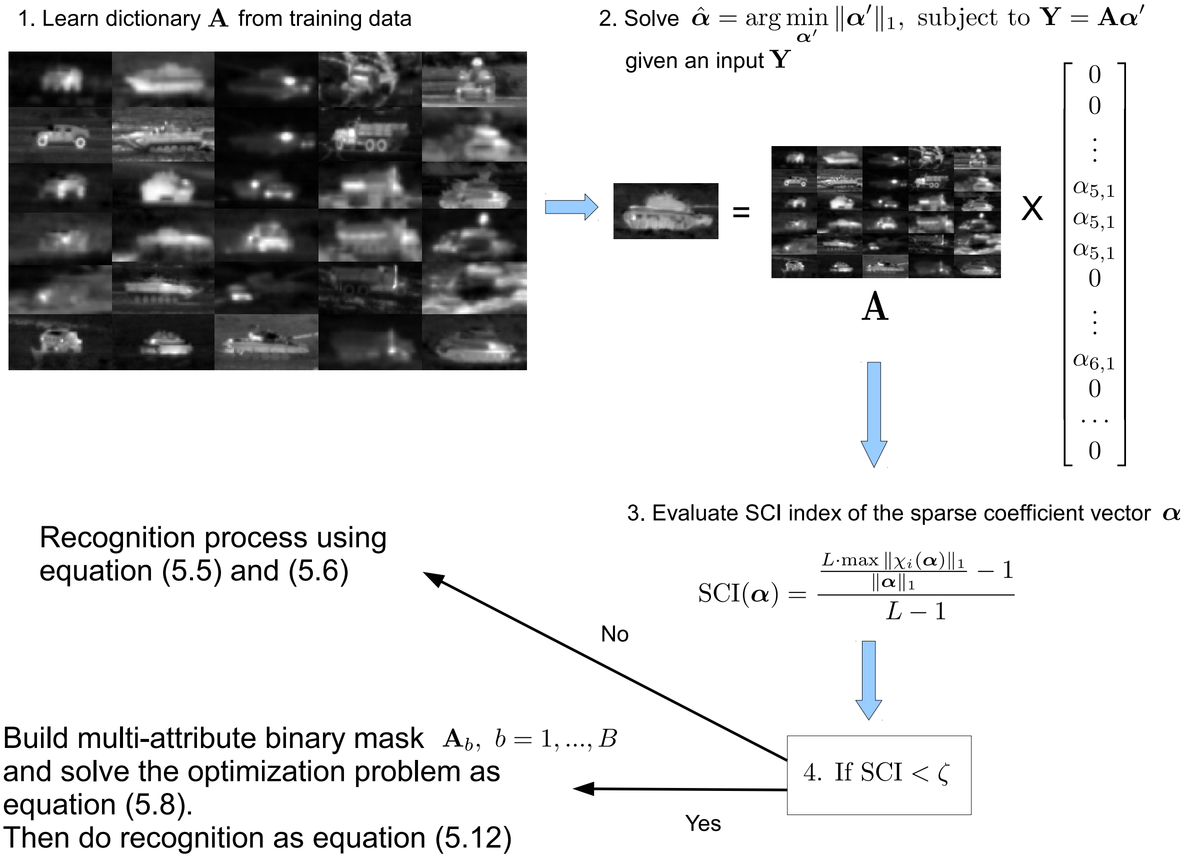

1. Learn dictionary $\mathbf{A}$ from training data
2. Solve $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}'} \|\boldsymbol{\alpha}'\|_1,$ subject to $\mathbf{Y} = \mathbf{A}\boldsymbol{\alpha}'$ given an input $\mathbf{Y}$
=
$\mathbf{A}$
X
$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ \alpha_{5,1} \\ \alpha_{5,1} \\ \alpha_{5,1} \\ 0 \\ \vdots \\ \alpha_{6,1} \\ 0 \\ \cdots \\ 0 \end{bmatrix}$
3. Evaluate SCI index of the sparse coefficient vector $\boldsymbol{\alpha}$
$\mathrm{SCI}(\boldsymbol{\alpha}) = \frac{\frac{L\cdot\max\|\chi_i(\boldsymbol{\alpha})\|_1}{\|\boldsymbol{\alpha}\|_1} - 1}{L-1}$
Recognition process using equation (5.5) and (5.6)
No
Build multi-attribute binary mask $\mathbf{A}_b,\ b = 1, ..., B$ and solve the optimization problem as equation (5.8).
Then do recognition as equation (5.12)
4. If SCI $< \zeta$
Yes

Figure 4.3: Flowchart of the swithable sparse representation classification (SSRC) method.

The coefficient vector $\boldsymbol{\alpha}$ is expected to be sparse enough so that training data associated to one class out of others would stand out for final recognition decision making. A sparsity concentration index (SCI) [103] was introduced to evaluate the quality of the optimization solution,

$$\mathrm{SCI}(\boldsymbol{\alpha}) = \frac{\frac{L\cdot\max\|\chi_i(\boldsymbol{\alpha})\|_1}{\|\boldsymbol{\alpha}\|_1} - 1}{L-1} \tag{4.14}$$

where $L$ is the number of class. SCI is between 0 and 1. The closer to 1, the sparser optimization result is achieved, and the test image could be approximated reconstructed by one single class of the training data.

Based on our experiments, although the multi-attribute method is approved to be effective on human face data [11], performances are very similar to the original SRC method on our Comanche dataset. One possible reason is that although the multi-attribute group constraint might encourage the optimization to be sparser, wrong classes which is similar to the correct one would stand out due to the group constraint. And also we found that most failure cases of the original SRC method occurs when the coefficient vector $\boldsymbol{\alpha}$ is not sparse enough. We also developed a third method which combines both of the original SRC and the multi-attribute SRC into a switchable sparse representation classification framework (SSRC) controlled by the SCI index. That is, if the SRC does not return a sparse enough solution - if the SCI index is lower than a threshold $\zeta$, the MALGC would get involved (Figure 4.3).

## 4.4 Experiments

In this section we will focus on a more challenging data - Comanche dataset, where target boundaries are less obvious and surrounding backgrounds are much more complex.

*(i) Comanche Data*

We will briefly introduce a much more challenging dataset for the ATR research - the Comanche dataset, as shown in Figure 4.4. This dataset contains ten targets (2S1 tank, M60 tank, M113 APC, M3 tank, M1 tank, Hummer SUV, BMP APC, T72 tank, M35 military truck and ZSU23 anti-aircraft tank), and there are 72 orientations for each target (0°,5°,10°, ... , 355°), and all targets are analyzed at 2000 meters in both day time and night time under various conditions such as different background, weather, in and around clutter. The data consists of SIG and ROI data, but we only

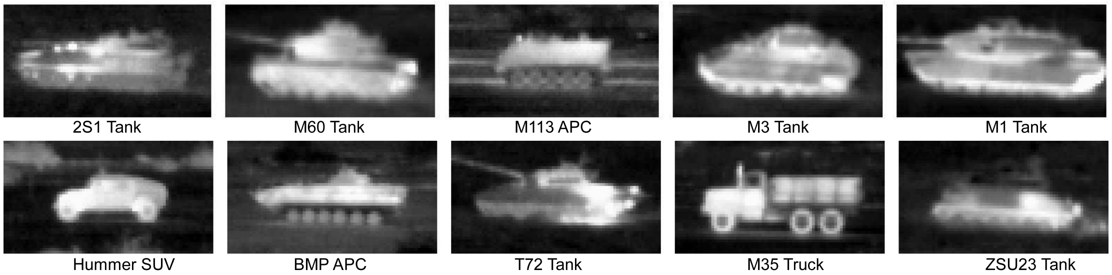


Figure 4.4: All ten targets in the Comanche dataset.

have the SIG data available, which has totally 13560 small image chips where targets are roughly located at the center of the image. We randomly select different groups of data for training, and we reduce the number of training data from 90% to 10% of the dataset each time, and the rest of the data are used for testing.

*(ii) Recognition Performances of SRC Methods*

Some data-driven methods - sparse representation based classification methods are applied in following experiments on the Comanche dataset. we will show some performances of both original sparse representation classification method from section 4.1.1 and the multi-attribute based method from section 4.1.2, and moreover, the results of the switchable sparse representation classification (SSRC) method.

With the SSRC method, the overall recognition accuracy could be much enhanced as shown in Table 4.1. In our experiments we found that the threshold of the SCI index is not sensitive, a better solution is setting it around 0.1 to 0.2 since we have 10 classes.

Table 4.1: Sparse representation-based target recognition accuracy (%) on Comanche dataset using different portion of data for training.

| **Training Data** | 90% | 80% | 70% | 60% | 50% | 40% | 30% | 20% | 10% |
|---|---|---|---|---|---|---|---|---|---|
| SRC | 95.68 | 94.96 | 94.35 | 93.12 | 92.15 | 91.32 | 87.63 | 82.14 | 74.12 |
| MALGC | 95.71 | 95.12 | 94.31 | 93.22 | 92.82 | 91.56 | 88.12 | 81.93 | 73.46 |
| SSRC | 95.92 | 95.31 | 94.75 | 93.92 | 93.76 | 93.13 | 89.83 | 83.92 | 75.42 |

## 4.5 Discussion

In this chapter we present a fusion technique for ATR tasks, where sparsity concentration index is used to combine two sparse representation classification methods, one is SRC and the other one is MALGC, during which we are able to add more group constraints from different attributes as simply solving a typical $l_1$ solution by reformulating the optimization problem into the matrix form. Experimental results shows promising benefit of the fusion framework than using only single sparse representation method. However, with the multi-attribute constraint, many testing data that were correctly classified by the original SRC could be mis-classified by adding group constraints, even the final results are slightly better. So in our future work we might try to find a more effective design of the SRC method that keep the group constraint in the optimization problem to encourage a more sparse solution of the coefficient vectors without sacrificing those correctly classified ones by the original SRC. Moreover, in recent years deep learning methods [159–161] become very popular for pattern recognition and machine learning researches, which is a next generation of computer vision research and are demonstrated to be very effective. So in the future we might explore deep learning techniques on ATR applications especially on FLIR data, to further develop an ATR system with improved efficiency and robustness.

# CHAPTER V

# RARE CLASSES ORIENTED IMAGE PARSING

Deep learning has been demonstrated to produce efficient and discriminative higher level features in image parsing problems in [75,84], where both features and classifiers are parameterized to model class based likelihood of pixels or superpixels so that different pixels in a given query image could be discriminated satisfactorily. But since CNN itself only captures object features in a local manner which does not contain any contextual information from the global scene, pure CNN methods may not work well in those pixels that look similar in a close and local view. For example the upper part of a cruise ship might look like a building, and sometimes road pixels might seems like sand or field without knowing the global scene information (road pixels appear more likely in the street scene rather than an image taken in the coast scene). To address this issue, an integrated non-parametric label transfer method was proposed in [13] into a parametric CNN framework, which utilize the global contextual information to alleviate the local ambiguity by using learned CNN features instead of human engineered ones.

However, in common computer vision applications, objects of interest typically occupy only small regions in the natural scene image or FLIR data, for example most ATR systems focus on humman, man made vehicles or structures in the observation rather than the background area. Compared to large background areas in the natural scene, objects with insufficient appearing frequencies and limited coverage of the whole image are considered rare class objects, which intensifies the difficulty of learning a robust recognition system. Although image parsing is able to assist specific object detection and recognition by fully using the surrounding background and contextual information, it is of great interests to extend the capability of a system to recognize rare class objects, such as human, birds, boats, fences, etc. With the idea of expanding the rare class subset [93] during image retrieval, we mitigate the work proposed in [13] for a better classification on rare classes without losing global understanding of query images, and an additional structure of rare class balanced CNN may further extend the ability of recognizing rare class objects. Specifically, our work makes following contributions,

- We selectively add rare class pixels to expand the retrieved image exemplar subset by using scene information, rather than equally selecting all rare classes. Through this we are able to add sufficient rare class objects into the feature space for label-transfer with lower computational cost and less class ambiguities.
- A complementary rare classes balanced CNN structure is trained to mitigate the negative effect caused by imbalanced training data problems. This extra CNN is implemented near the region of rare classes upon the combination of the original CNN local belief and the global class likelihood by label transfer. Reduced effect on frequent classes and less computational complexity can be achieved through this additional CNN structure for better labeling results.
- A superpixels-based re-segmentation is adopted to enhance perceptually meaningful object boundaries by taking advantage of CNN-based pixel labeling.

## 5.1 Preliminary Work

In this section, we will first introduce some background on deep learning and convolutional neural networks, followed by a brief introduction on a hybrid scene labeling framework which integrates a parametric CNN structure for local labeling and a non-parametric label transfer process for global labeling.

### 5.1.1 Deep Learning and Convolutional Neural Networks (CNNs)

For decades, great efforts have been made to duplicate the functioning procedures of human brains at least partially in artificial intelligence research [162]. Deep learning is intuitively motivated by studies on biological nature of human brains which showed that the neocortex in our brain process data through a propagation of some complex hierarchy modules to learn observations [163]. Such hierarchical networks make it possible to train deep artificial networks on a large-scale dataset with simple classifier for robust pattern recognition tasks.

Deep learning [159–161] has been widely used in the field of pattern recognition and machine learning in recent years. It is a group of methods that build a deep architecture with many layers of adaptive non-linear components, which are cascades of parameterized non-linear modules that contain trainable parameters at all levels [164]. With deep learning algorithms, not only can we find a way to specify prior knowledge in a flexible way using deep architectures that can handle large families of functions for training, but also use this concept for multi-task learning and semi-supervised learning [159]. The most popular deep architectures are convolutional neural networks (CNNs) [77] and deep belief network (DBN) [165], stacked autoencoders [166], restricted Boltzmann machine (RBM) [167], etc. And furthermore, unsupervised feature pre-training is suggested in [168] before the supervised learning process for better learning results.

CNNs have played significantly important roles in many computer vision applications, such as hand-written digit classification [76], object recognition [169], robotic vision [80], segmentation [81], detection [82], scene labeling [13], and so on. CNNs have the capability of building translational invariant features through pooling, and also have significant advantages of fewer parameters [170]. As the on-going advancement of computational hardware power, collecting a large-scale labeled dataset becomes much easier and models with large learning capability with sufficient prior knowledge incorporated are highly encouraged, and CNNs have been demonstrated to be extremely useful [77, 171].

CNNs involve discrete convolution on spatial information between pixels of images or videos [78], and they are among the first truly successful deep learning approach where many layers of a hierarchy were successfully trained in a robust manner. A CNN is simply a framework of topology or architecture that provides locally connected (spatially and temporally) structure which is capable of reducing the model complexity, and optimization process is simplified by minimizing the number of controllable parameters. Usually such parameters must be learned upon general feed-forward back propagation. One important motivation that triggers the development of CNNs was to build a deep learning framework with less pre-processing. Thus, only small portions of the input are involved in the local receptive fields into the lowest layer of the hierarchical structure [162].

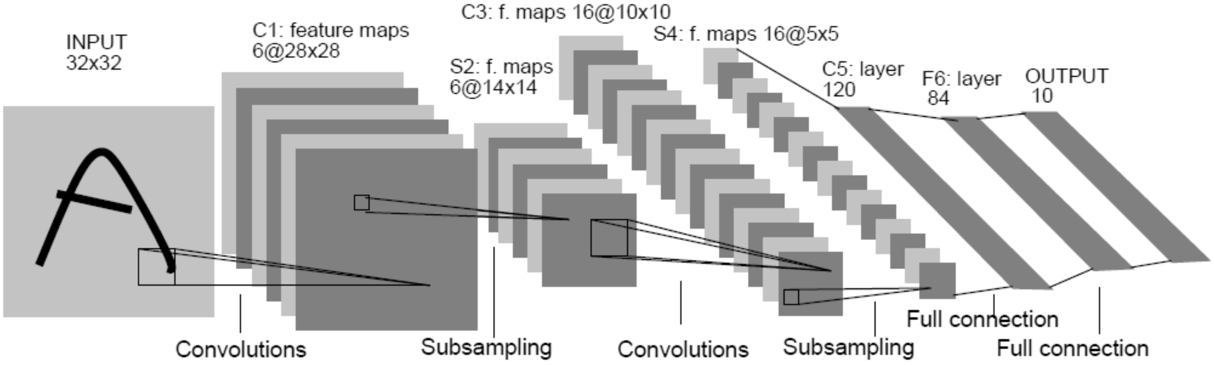


Figure 5.1: LeNet-5 deep convulitional neural network architecture for digit recognition [12].

CNN is designed to benefit from the 2D structure of observations, which is achieved by locally connected neurons tied with weights and biases followed by pooling layers in the purpose of translation invariance. A typical CNN consists of convolutional layers, spatial pooling layers and full-connected layers. Suppose the input to the first convolution layer have the dimensionality of $n \times n \times r$, where $n$ is the height and width of the input image and $r$ is the number of channels. For example as shown in Figure 5.1, the input is simply a $32 \times 32$ gray scale image with $r = 1$. The convolution layers consist of $q$ filters of $m \times m \times k$, where $0 < k <= r$, $m$ is the size of filters, and $m$ is smaller than $n$ which results in the locally connected architectures in purposes of the convolution, with the output of $q$ features with the size of $n - m + 1$. After the convolution layer, typically each feature map is subsampled with selected pooling methods, such as mean or max pooling [172].

One could use all features produced from convolution layers for classification, but there are two advantages by incorporating the pooling layers into the deep architecture. The first one is dimension reduction, since the convolution usually results in features with a large size of feature maps, and it is difficult to use all of these features for future convolution and classification. The second advantage is that pooling could result in statistical aggregation in the 2D input. More specifically, since 2D images stay stationary and features that are able to describe one region of the image might be applied to some other regions as well, and such benefit could be desirable for us to build hierarchical networks that are robust to translation invariance [173].

Then, we might have some fully connected layers which are similar to the one in standard multilayer neural networks. In classification problems for example, one might need to add some type of classifier to output a posterior probability. Here we will briefly introduce one typically used model named softmax regression [174]. Suppose we are given a labeled training dataset $\{(x^1, y^1), \cdots, (x^m, y^m)\}$ , where $x^i$ is the $i_{th}$ training image with label $y^i \in \{1, 2, \cdots, L\}$. Thus, we try to give a hypothesis or belief of an input image with a probability that $P(y = l|x)$, where $l = 1, \cdots, L$ are labels of different classes. Hence we want to obtain a $L \times 1$ vector where each element gives a probability of the input for each different classes, and the hypothesis is in the form of

$$p_\theta(x) = \begin{bmatrix} P(y=1|x;\theta) \\ P(y=2|x;\theta) \\ \vdots \\ P(y=L|x;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{L} \exp(\theta_j^T x)} \begin{bmatrix} \exp(\theta_1^T x) \\ \exp(\theta_2^T x) \\ \vdots \\ \exp(\theta_L^T x) \end{bmatrix}, \tag{5.1}$$

where $\theta$s are the parameters of our model. And a typical cost function for the softmax regression can be described below,

$$J(\theta) = -\left[ \sum_{i=1}^{m} \sum_{k=1}^{L} \delta(y^i = k) \log \frac{\exp(\theta_k^T x^i)}{\sum_{j=1}^{L} \exp(\theta_j^T x^i)} \right], \tag{5.2}$$

where $\delta(\cdot)$ is a indicator function ($\delta(true) = 1, \delta(false) = 0$). This softmax cost function is usually hard to be minimized analytically, and iterative optimization algorithms can be used [172].

### 5.1.2 Integrated Parametric and Non-parametric Models

CNNs have been effectively used in scene labeling problems in [75, 84], where parameterized CNN models were proposed to learn pixel-wise class likelihood in a local context-based manner. But pure CNN models are not able to handle global contextual information as various features in different classes might be similar in a close and local view at the pixel level, and some objects have less possibility to appear under certain circumstances as well. For example, pixels of the sand may look very similar to those of the road; it is less likely to observe a car running in the ocean; and the sky will most probably appear above the land or the sea in the upper part of an image. In [13] an integration of parametric and non-parametric models was proposed to address this issue, where not only CNN models were adopted for feature learning and local classification, but also a non-parametric label transfer method was incorporated to achieve enhanced classification robustness for locally indistinguishable pixels.

This integration have three advantages. First, the global semantics helps to remove pixel level ambiguities by providing class dependencies globally from nearest exemplars in the training dataset. Second, discriminative features learned from CNN were used for global feature learning instead of human engineered low-level features. Third, a relatively small retrieval sets are sufficient due to the global scene constraint. Specifically, the overall objective function is defined as follows,

$$E(\mathbf{X}, \mathbf{Y}) = -\sum_{i \in \mathbf{X}} (P_L(X_i, Y_i) + P_G(X_i, Y_j)) \tag{5.3}$$

where $\mathbf{X}$ is the observation image, $X_i$ is the $i_{th}$ pixel and $Y_j$ is the label corresponding to $X_i$, and $j = 1, 2, \cdots, L$ and $L$ is the number of classes associated to pixels. $P_L(X_i, Y_i)$ is the local belief for each pixel obtained by training a parametric CNN model on small image patches as shown in Figure 5.2, and $P_G(X_i, Y_i)$ is the non-parametric global belief achieved by a weighted K-nearest neighbor (KNN) in a global features space obtained by the trained CNN.

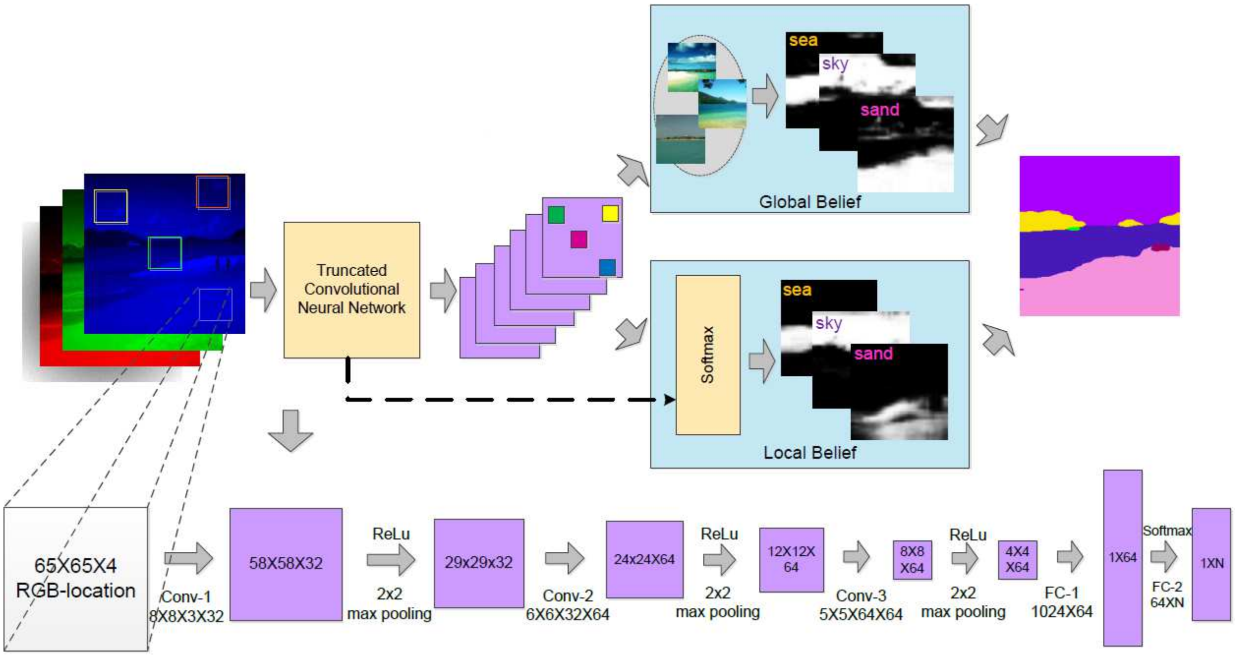


Figure 5.2: Parametric and non-parametric integrated scene labeling framework [13].

Specifically, after the training process of the CNN, the corresponding CNN feature tensors $\mathbf{F} \in \mathbb{R}^{H \times W \times T}$ could be obtained by passing the input images to the truncated CNN (i.e. the CNN without the softmax layer), where $H$ and $W$ are the height and width of the image, and $T$ is defined by the last layer output of the truncated CNN. For simplicity an image could be decomposed into several region blobs $\mathbf{B} = \{\mathbf{B}_1, \mathbf{B}_2, ..., \mathbf{B}_n\}$, thus the region feature is defined as $\mathcal{H}(\mathbf{B}_i) = pool(\mathbf{F}^i), \forall i \in \mathbf{B}_i$ by average pooling [77] from the constituent features. Moreover, a global feature space $\mathcal{H}$ is defined by including all feature tensors $\mathcal{H}(\mathbf{B}_i)$ in the training dataset. The pixel-wise global belief is defined as

$$P_G(X_i, Y_j) = \frac{\sum_k \Phi(X_i, X_k)\delta(Y(X_k) = Y_j)}{\sum_k D(X_i, X_k)}, \quad \forall X_k \in \mathcal{S}(X) \tag{5.4}$$

where $X_k$ is the $k_{th}$ nearest neighbor of pixel $X_i$ within the nearest exemplars $\mathcal{S}(\mathbf{X})$ defined in the global feature space $\mathcal{H}$ from all the training data, and $Y(X_k)$ is the ground truth label of the pixel $X_k$, and $\delta(\cdot)$ is equal to 1 if $Y(X_k) = Y_j, \forall j = 1, 2, ..., L$, and $L$ is the number of classes. $D(\cdot)$ gives the distance or similarity between the query pixel and the training pixel from nearest exemplars in the CNN feature space,

$$D(X_i, X_j) = \exp(-\alpha\|x_i - x_j\|)\exp(-\gamma\|z_i - z_j\|) \tag{5.5}$$

where $x_i$ is the CNN feature corresponding to pixel $X_i$. $z_i$ is the coordinate along the vertical direction of the image. $\alpha$ and $\gamma$ control the trade-off between the feature and spatial distances.

## 5.2 Scene Assisted Rare Classes Retrieval

As shown in Figure 5.3, usually objects in natural scene images are not equally distributed, because pixels of rare classes (human, boats, windows, cars, etc.) have low appearing frequency and small spatial coverage compared to common classes (sky, buildings, trees, etc.). A *hybrid* sampling method was introduced during CNN training in [13], where pixels in rare classes were manually given higher frequency to appear in the sampled training patches. Yang et al. mitigated the imbalanced data distribution problem by simply adding superpixels from rare classes to the retrieval

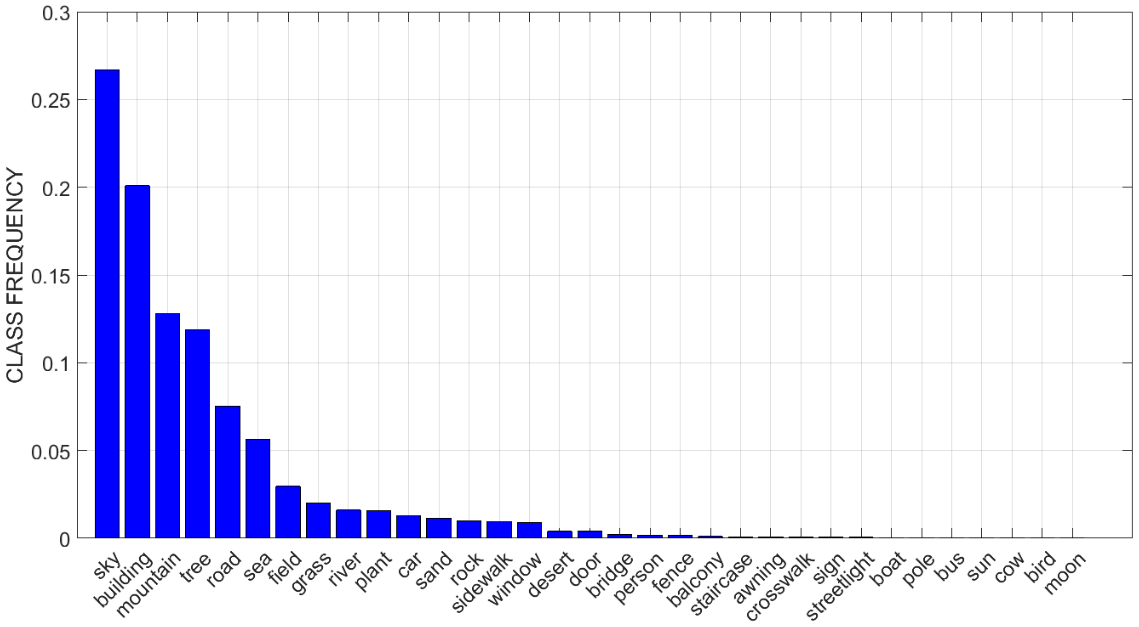


Figure 5.3: Unbalanced class distribution in the SIFTflow dataset [1].

set, which expands the retrieved subset to have more balanced frequency distribution among all classes [93]. This method is able to increase the number of rare classe superpixels in the retrieval set, leading to a more balanced label transfer. But for each query image, the expanded subset is obtained by random sampling among all rare classes, which does not encourage global contextual information. Hence, we fuse a scene assisted rare classes retrieval method into the non-parametric model [13] for global label transfer, where scene information is incorporated into the expansion of the rare class subset from the training dataset for enhanced labeling performances on rare class objects.

To demonstrate the importance of scene information during rare class retrieval, Table 5.1 summarizes some frequencies of selected classes in the SIFTFlow dataset [1] according to the scene layouts (coast, forest, highway, city, mountain, country, street, and tall buildings as shown in Figure 5.4). As shown in the table, frequencies are closely related to the image scene information, for example humans appear most probably in the street scene, sidewalk has the largest frequency in the street scene,

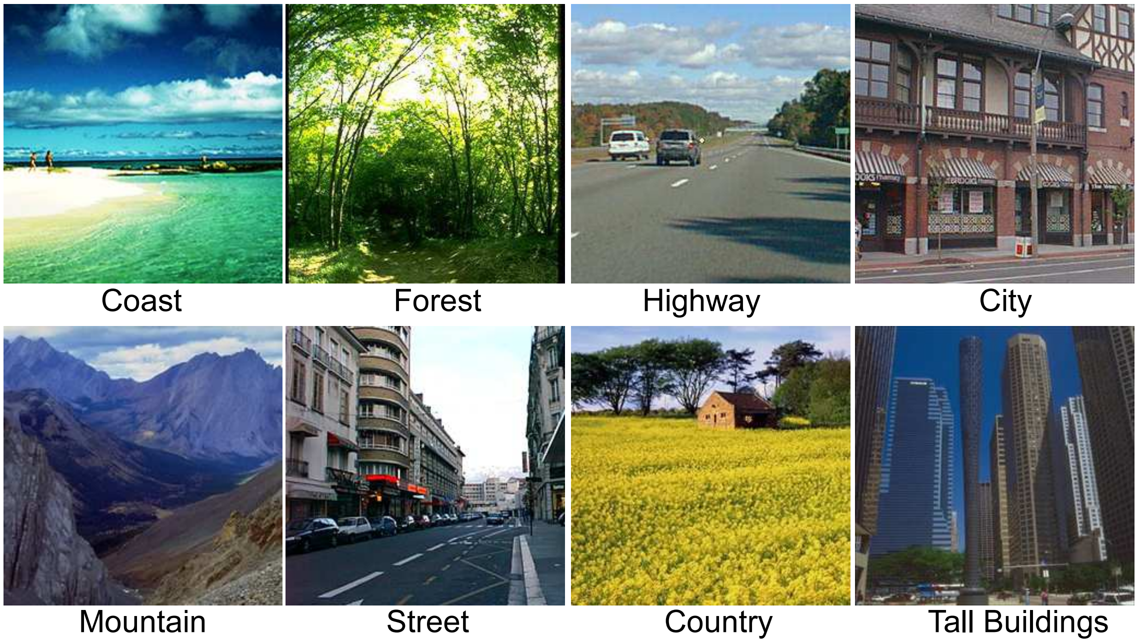


Figure 5.4: Natural scenes in the SIFTflow dataset [1].

Table 5.1: Selected class frequency (%) of scene categories in the SIFTflow dataset [1]. (Hwy=highway, Mnt=mountain, St=street, Bldg=building).

| | Coast | Forest | Hwy | City | Mnt | Country | St | Tall Bldg |
|---|---|---|---|---|---|---|---|---|
| **Bldg** | 0.38 | 0.17 | 2.32 | 58.17 | 0.18 | 0.52 | 40.91 | **60.03** |
| **Car** | $< 0.01$ | 0.02 | 2.37 | 2.44 | $< 0.01$ | $< 0.01$ | **6.26** | 0.28 |
| **Person** | 0.04 | 0.08 | 0.01 | 0.31 | 0.09 | 0.02 | **0.94** | 0.06 |
| **Boat** | **0.13** | $< 0.01$ | 0 | 0.03 | 0 | 0.02 | 0.04 | 0.06 |
| **Window** | 0 | 0 | $< 0.01$ | **6.77** | $< 0.01$ | $< 0.01$ | 0.45 | 0.15 |
| **Sidewalk** | 0 | 0 | 0.44 | 3.24 | 0 | 0 | **4.32** | 0.25 |
| **Road** | 0.04 | 0.19 | **36.08** | 8.56 | 0.23 | 0.51 | 24.52 | 0.76 |
| **Field** | 0 | 0.07 | 1.42 | 0 | 0.16 | **19.05** | 0 | 0 |
| **Balcony** | 0 | 0 | 0 | 0.8593 | 0 | 0 | 0.0298 | 0 |

and cars are more likely to appear in the street scene rather than in the forest scene, while field pixels have higher chance to appear in the countryside rather than in the city. And also we can see clearly that, it is almost zero or less than 0.01% for boats to appear in the mountain, forest or highway (some trucks might be towing a boat on the highway though), and balcony appears only in the city and street scenes. Encouraging the image scene information during the retrieval process may help to preserve rare classes more accurately with lower ambiguities among rare class objects.

The purpose of image retrieval is to find image exemplars from the training dataset with similar scene layout to the input, so that label transfer can be achieved by scene constrained object labels. But rare class objects in the test image may not appear in the retrieved image set, since only global scene layout is preserved. In this case, misclassification may occur due to the fact that rare class objects are possibly missing, So it is important to enrich the retrieval set with rare classes. However, random selection [93] brings all rare classes to the retrieval set, which may not be relevant to the scene category. For example, in a highway scene, it is not necessary to consider adding boat pixels, as shown in Table 5.1, frequency of boat in a highway scene is almost close to zero. Although random selection seems effective when the retrieval fails (i.e. retrieved images are from different scene categories compared with the input), it is important to note that retrieval is a significant guarantee for accurate label transfer, and failed retrieval may result in misclassification for all classes. In order to incorporate the scene information into the retrieval process, we consider a $L_r \times 1$ frequency vector $\mathbf{f}_c$ for each scene category $c \in \{1, 2, \cdots, C\}$, where $C$ is the number of scene categories ( $C = 8$ in the SIFTflow dataset). Each element of $\mathbf{f}_c$ is the frequency of each rare class in the scene category $c$, and $L_r$ is the number of rare classes. This frequency vector will encourage sampling of rare classes constrained by scene information. For scene assisted sampling, we obtain the sampling ratio vector

$\mathbf{s}_c$ as follows,

$$\mathbf{s}_c = \frac{\mathbf{f}_c}{\sum \mathbf{f}_c} \tag{5.6}$$

where $\mathbf{s}_c$ provides the proportion of each rare class in the resulting retrieval set of rare classes under the scene category $c$.

During the retrieval step, suppose we have a trained CNN model $\mathcal{M}$ from the training dataset $\mathbf{X} = [\mathbf{X}_1,\ \mathbf{X}_2, \cdots,\ \mathbf{X}_N]$ with labels $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_N]$, $\mathbf{Y} \in \{1, 2, \cdots, L\}^N$, where $N$ is the number of training images and $L$ is the number of classes, we could obtain feature tensors $\mathbf{F} \in \mathbb{R}^{H \times W \times T}$ from each images by applying the CNN model to each image, where $H$ and $W$ are the height and width of the image, and $T$ is defined by the last layer output of the truncated CNN.

Given a query image $\mathbf{X}_q$, similarly we can obtain a retrieval set $\mathcal{S}_q(\mathbf{X})$ with $m$ exemplars from the training dataset by CNN feature matching, and hence the scene category $c$ could be obtained based on the dominant scene in $\mathcal{S}_q(\mathbf{X})$. We introduce a scene assisted retrieval set $\mathcal{B}_q(\mathbf{X})$ for rare class objects, and the proportion of each rare class in $\mathcal{B}_q(\mathbf{X})$ follows $\mathbf{s}_c$. The number of samples $S^l$ needed for each rare class $l$ is given by

$$S^l = \begin{cases} M_{\mathcal{S}_q(\mathbf{X})} \cdot \mathbf{s}_c^l & \\ N_l & \text{if} \quad N_l < M_{\mathcal{S}_q(\mathbf{X})} \cdot \mathbf{s}_c^l \end{cases} \tag{5.7}$$

where $M_{\mathcal{S}_q(\mathbf{X})}$ is the average number of pixels among dominant classes (such as sky, sea) in $\mathcal{S}_q(\mathbf{X})$, which ensure a balanced distribution of all classes (both frequent classes and scene-relevant rare classes). $N_l$ is the total number of pixels belong to rare class $l$ all over the training dataset. By sampling $S^l$ number of pixels for each rare class $l$ from the training dataset, we are able to obtain an expanded retrieval exemplar set $\mathcal{Z}_q(\mathbf{X}) = \mathcal{S}_q(\mathbf{X}) \cup \mathcal{B}_q(\mathbf{X})$. This scene assisted rare classes retrieval is summarized in Algorithm 4.

The global belief for a given query image could be obtained by transferring labels

**Algorithm 4** Scene Assisted Rare Classes Retrieval

```
CNN model $\mathcal{M}$ learned from data with natural distribution,
    training images $[\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N]$. Ground Truth Labels $[\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_N]$.
Input all Images $\mathbf{X}_{1:N}$ to $\mathcal{M}$, acquire feature tensors $[\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_N]$.
Obtain global features $[\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_N]$ by average pooling.
Based on the global image category $c \in \{1, 2, \cdots, C\}$ for all training images, compute $C \times L_r$
    frequency matrix $\mathbf{f}$ for $L_r$ rare classes, $L_r < L$ of each image category $c$.
Given a query Image $\mathbf{X}_q$, similarly obtain $\mathbf{H}_q$, retrieve nearest exemplars $\mathcal{S}_q(\mathbf{X})$ from training
    images by matching $\mathbf{H}_q$ and $\mathbf{H}$.
Assign the image category label $c$ for the query image based on $\mathcal{S}_q(\mathbf{X})$, and then Obtain $\mathbf{f}_c$ of
    rare classes from $\mathbf{f}$, and compute $\mathbf{s}_c$ by Equation (5.6).
Sample rare class image patches according to Equation (5.7) to obtain $\mathcal{B}_q(\mathbf{X})$.
Final retrieval set for query image $\mathbf{X}_q$ is $\mathcal{Z}_q(\mathbf{X}) = \mathcal{S}_q(\mathbf{X}) \cup \mathcal{B}_q(\mathbf{X})$
```

of the retrieval set $\mathcal{Z}_q(\mathbf{X})$ by the weighted KNN equation as,

$$P_G(X_i, Y_j) = \frac{\sum_k D(X_i, X_k)\delta(Y(X_k) = Y_j)}{\sum_k D(X_i, X_k)}, \quad \forall X_k \in \mathcal{Z}_q(\mathbf{X}) \tag{5.8}$$

where, $X_k$ is the $k_{th}$ nearest neighbor of the pixel $X_i$ within the expanded nearest exemplars $\mathcal{Z}_q(\mathbf{X})$ defined in the global feature space from all training data, and $Y(X_k)$ is the ground truth label of pixel $X_k$, and $\delta(\cdot)$ is equal to 1 if $Y(X_k) = Y_j, \forall j = 1, 2, ..., L$, and $L$ is the number of classes. $D(\cdot)$ gives the distance or similarity between the query pixel and the training pixel in the nearest exemplars in the feature space constrained by spatial information,

$$D(X_i, X_j) = \exp(-\alpha\|x_i - x_j\|)\exp(-\gamma\|z_i - z_j\|)\exp(-\beta\|b_i - b_j\|) \tag{5.9}$$

where $x_i$ is the CNN feature corresponding to pixel $X_i$; $z_i$ is the coordinate along the vertical direction of the image; $b_i$ is the object size of the image blob where the pixel $X_i$ belong to, which ensure that we evaluate the distance of two features not only by feature similarities, but also by the pixel location and the object size. This distance function encourage a larger dissimilarity between rare classes and frequent classes.

For example in outdoor scenes, the sky may appear above the sea, and a person is likely to appear in the lower part of an image, while the region that is occupied by a human or a boat is usually smaller compared to the sky or the sea. $\alpha$, $\gamma$ and $\beta$ control trade-off between the CNN feature, the spatial distance and the object size.

## 5.3 Rare Classes Balanced CNN

Since the local belief was estimated by training a CNN using small image patches around each pixel, it is hard to be adapted to all classes when pixel distributions are not well balanced, and objects with smaller areas covered in the image are tend to be dominated by frequent class objects. In other words, sufficient training may be processed for frequent classes, such as sky, sea, sand, but training image patches available for rare classes are insufficient, such as person, boat, car, etc. So in this work we propose a rare classes-oriented scene labeling framework (RCSL), which adds an extra structure of CNN trained with a subset of the training data with balanced distribution among all classes, during which enough learning process could be implemented for better feature learning. Specifically, we develop an energy function $E_r(X_i, Y_j)$ that evaluates the class likelihood of the label $Y_j, j = \{1, \cdots, L\}$ only for rare classes pixels $X_i$ in an image $\mathbf{X}$,

$$E_r(X_i, Y_j) = -P_{Lr}(X_i, Y_j) \cdot \delta(Y(X_i) \in \mathcal{Y}_r), \tag{5.10}$$

where $\mathcal{Y}_r$ is a subset of labels $\mathcal{Y} = \{1, 2, \cdots, L\}$ which only contains rare class labels. $\delta(\cdot)$ is an indicator function and $\delta(Y(X_i) \in \mathcal{Y}_r) = 1$ if $Y(X_i) \in \mathcal{Y}_r$, and

$$Y(X_i) = \arg \min_{1,\cdots,|L|} E(X_i, Y_j), \tag{5.11}$$

where

$$E(X_i, Y_j) = -(P_L(X_i, Y_i) + P_G(X_i, Y_j)), \tag{5.12}$$

where $E(X_i, Y_j)$ is the class likelihood obtained by the local belief $P_L(X_i, Y_i)$ achieved by a CNN model $\mathcal{M}$ [13], and the global belief $P_G(X_i, Y_i)$ from Equation (5.8). $P_{Lr}$ is computed by training a separate CNN model $\mathcal{M}_r$ based on a more balanced training data over all classes, especially determined by the frequency distribution of rare classes. Since the training data contains significantly large training image patches, we are sampling [175] the whole training dataset so that only part of them is used to train $\mathcal{M}_r$. In this manner, we first randomly sample image patches of rare classes until they are equally distributed in the training patches set, and then based on the average occurrence rate of the rare class in the training subset, all frequent classes are sampled in order to reach the same occurrence rate in the final training set. specifically, suppose we are sampling $N_s$ pixels for each rare class, in which case, patches from the rare class whose total number of pixels is less than $N_s$ may appear multiple times, and only $N_s$ Pixels from frequent classes are sampled to keep the training data balanced over all classes.

**Algorithm 5** Rare Class-oriented Scene Labeling (RCSL)

```
training images $[\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N]$. Ground Truth Labels $[\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_N]$.
Train a CNN-Softmax $\mathcal{M}$ (natural distribution) and $\mathcal{M}_r$ (rare class balanced distribution),
for each testing image $\mathbf{X}_q$, $q = 1, \cdots, Q$ do
    Find image retrieval set $\mathcal{Z}_q(\mathbf{X})$ as Algorithm 4,
    for pixel $i = 1, \cdots, H \times W$ do
        compute local belief $P_L(X_i, Y_j)$ using $\mathcal{M}$, global belief $P_L(X_i, Y_j)$ as Equation (5.8),
        obtain $Y(X_i)$ as Equation (5.11),
        calculate $E_r(X_i, Y_j)$ as Equation (5.10),
    end for
    obtain the labeling map $\mathbf{Y}_q$ of the input image $\mathbf{X}_q$ by Equation (5.13).
end for
```

Hence, the final RCSL labeling $Y$ can be accomplished in a pixel-wise manner since pixels are independent within each other,

$$\mathbf{Y} = \cup_{i=1:N} Y_i \tag{5.13}$$

where

$$Y_i = \arg \min_{1,\cdots,|L|} (E(X_i, Y_j) + E_r(X_i, Y_j)) \tag{5.14}$$

We summarize the final RCSL labeling with scene assisted rare class retrieval in Algorithm 5.

## 5.4 Superpixels-based Re-segmentation

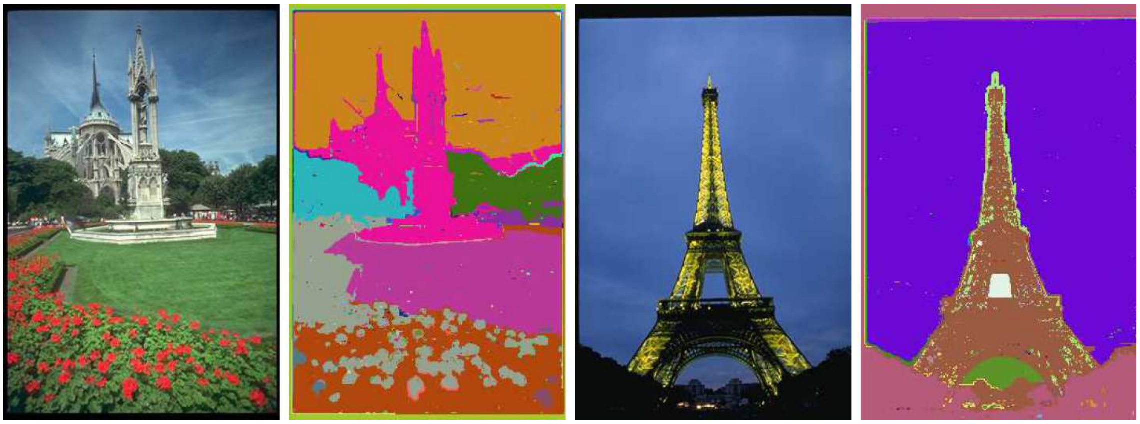


Figure 5.5: Graph-based image segmentation capturing perceptually important groupings or regions [14].

Image segmentation and grouping have been widely used in scene labeling problems, for example in [93], pre-segmentation of finding superpixels offers perceptually important neighboring regions or groupings which captures object boundaries effectively. Felzenszwalb et al. proposed an efficient graph-based image segmentation algorithm [14] (as shown in Figure 5.5) based on two important perceptual aspects: first, widely varying intensities are not supposed to be judged as the evidence to separate different regions; second, meaningful regions cannot be obtained using only local decision criteria. These two aspects lead to a segmentation algorithm that compares

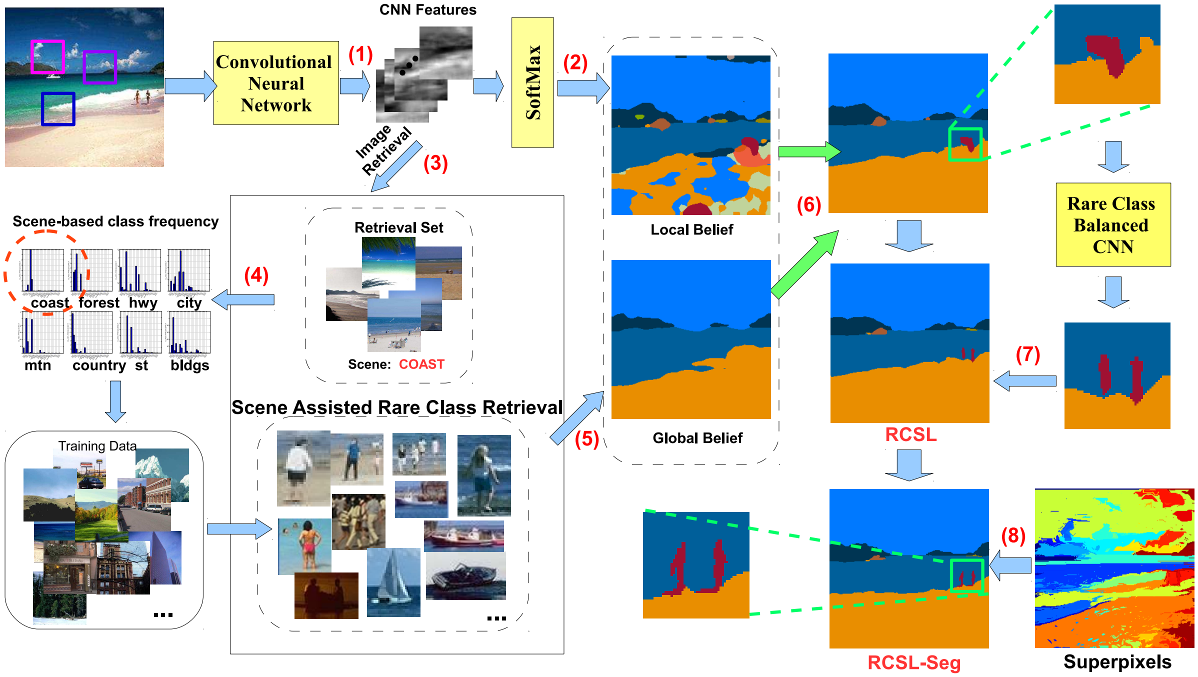


Figure 5.6: Proposed RCSL and RCSL-Seg framework [15]. (1) Pass the input image to obtain CNN features for each pixel; (2) Local belief computed by the original CNN [13]; (3) retrieve exemplars from the training dataset; (4) Scene assisted rare class retrieval by Algorithm 4; (5) compute the global belief as Equation (5.8); (6) find rare class objects by combining global and local belief; (7) obtain complementary local belief according to Equation (5.13); (8)superpixels-based post-processing according to Equation (5.15).

difference not only across boundaries, but also between neighboring pixels within each region.

Since the resulting segmentation from the proposed RCSL tends to be smeared and rough, where only image patches are used during CNN training without consideration of any boundary information. In order to alleviate this problem, we develop a superpixels-based re-segmentation method that is integrated into our RCSL (RCSL-Seg) for better segmentation and labeling accuracy.

Given an input image $\mathbf{X}$, we are able to obtain a group of superpixels $R_p$, $p =$

$1, 2, \cdots, P$ obtained by the graph-based segmentation [14], and these superpixels divide the image $\mathbf{X}$ into $P$ regions. Suppose each pixel $X_i \in R_p$ is labeled as $Y_i$ ($Y_i \in \{1, 2, \cdots, L\}$ and $L$ is the number of classes) based on the RCSL framework with scene assisted rare class retrieval, we further refine the final segmentation by evaluating the majority of labels in each superpixel $R_p$. The final label $Y'_p$ for each superpixel is given by

$$Y'_p = \arg \max_{j=1,\cdots,L} N_j \tag{5.15}$$

where $N_j$ is the number of pixels labeled as $j$ in the superpixel $R_p$. Thus, each superpixel could be labeled based on the RCSL framework for refined segmentation. The overall RCSL and RCSL-Seg system with scene assisted rare classes retrieval is shown in Figure 5.6.

## 5.5 Experiments

The SIFTflow dataset [1] has been widely evaluated in may scene labeling work [13, 93, 96], which consists of 2688 images of size $256 \times 256$ from eight scenes containing 33 semantic class objects captured under eight typical natural scenes, including coast, forest, highway, city, mountain, street, country and tall buildings (Figure 5.4). All class labels and frequencies in the whole dataset are shown in Figure 5.3. The dataset has been separated into 2488 training images and 200 testing images equally distributed in each natural scene. In the dataset, each image is associated with ground truth segmentation with corresponding labeling to each pixel, as shown in Figure 5.7. In this section we conduct three studies to show the advantages of proposed rare-class enhanced scene labeling framework with refined segmentation results using superpixels. In the first study we will show the visual effectiveness of using superpixels to refine the final labling segmentation regions, which lead to more detailed object boundaries that are closer to those of real world objects. In the second study, we report the overall labeling accuracies of proposed methods and several re-

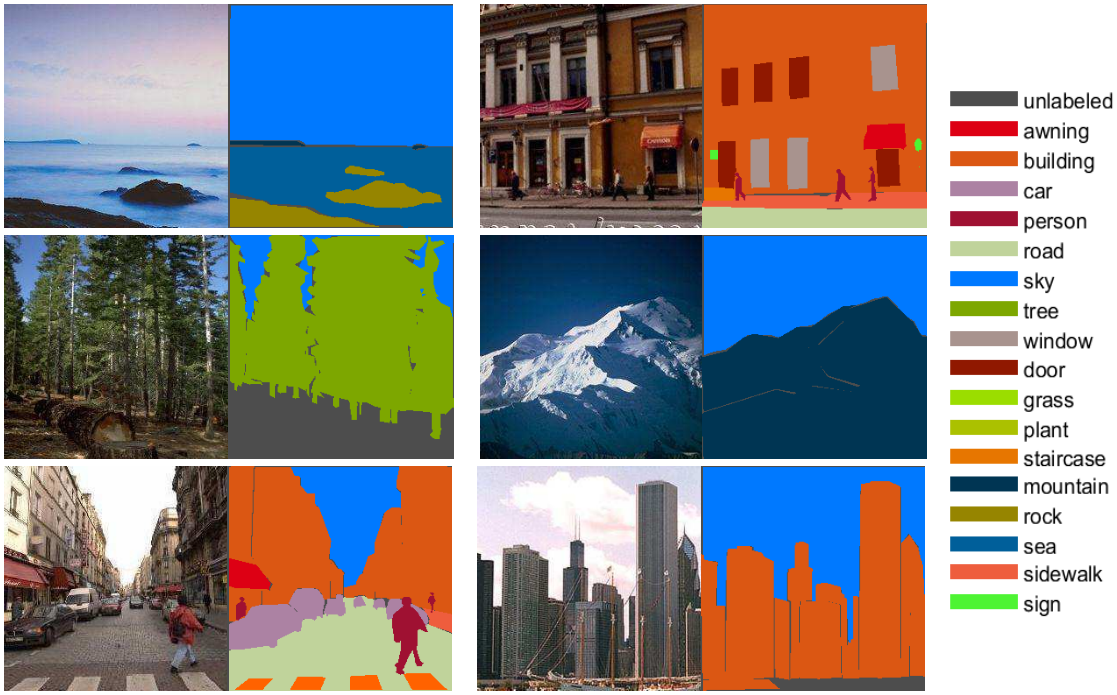


Figure 5.7: Examples from the SIFTflow dataset [1]. The 1st and 3rd columns are the color images from natural scenes, the 2nd and 4th columns are the corresponding ground truth segmentations and pixel-wise semantic labels.

cent research. Three versions of RCSL are evaluated, including RCSL-I(RCSL with random sampling among rare classes), RCSL-II (RCSL with scene assisted rare class retrieval), and RCSL-Seg (RCSL with superpixels-based re-segmentation) [14]. And finally, we evaluate both pixel and class accuracy for rare classes to demonstrate the promising performances of proposed methods quantitatively.

### 5.5.1 Superpixels-based Re-segmentation

The proposed RCSL method is basically using the combination of CNN for local labeling and nonparametric label transfer model for global labeling [13], where small image patches are used to train the CNN model. Our CNN models are trained on Nvidia Geforce GTX 660 using MatConvNet toolbox [176], which takes about 5 hours for training the local CNN and 3 hours for the rare classes enhanced CNN.

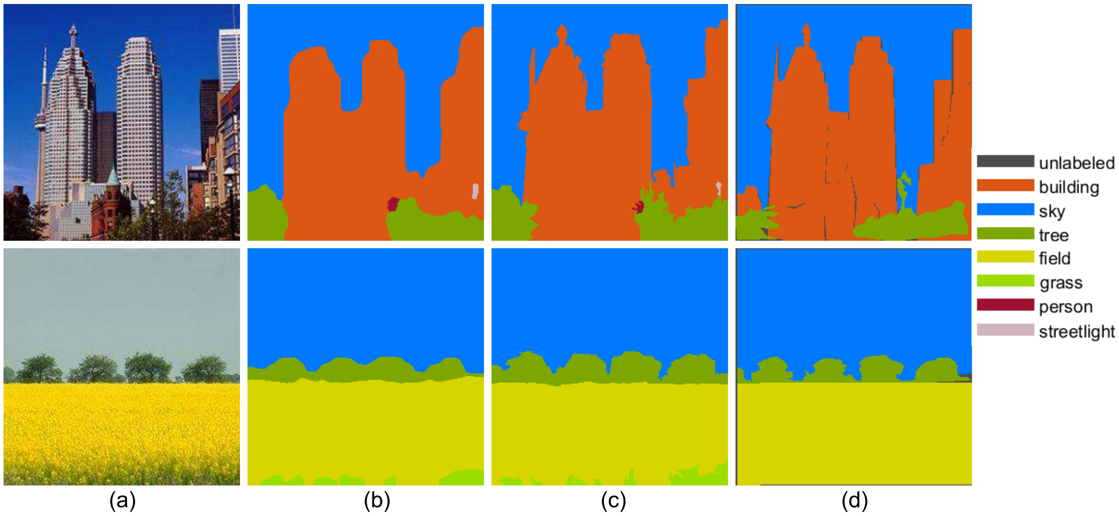


Figure 5.8: Demonstration of the re-segmentation using superpixels [14]. (a) The original image, (b) RCSL, (c) RCSL with superpixels-based re-segmentation, and (d) the ground truth.

Although the final labeling results tend to preserve the global semantic layouts, the resulting segmentation is found to be smeared since object boundaries or shapes are not considered. In order to obtain better segmentation and higher labeling accuracy, we implemented a simple yet efficient graph-based image segmentation [14] to refine the final labeling results obtained from the RCSL algorithm, where object boundaries are demonstrated to be close to real-world objects. As demonstrated in Figure 5.8, we can see clearly that boundaries of the building are well recovered after the superpixels refinement, and especially those trees tend to be more detailed and accurate. This refinement will not only refine the final segmentation masks but also contribute to better labeling accuracy.

### 5.5.2 Overall Labeling Accuracy

We report the quantitative results of proposed methods on the SIFTflow dataset in Table 5.2, including RCSL-I (rare class random sampling), RCSL-II (scene assisted rare class retrieval), RCSL-Seg (RCSL with superpixels-based re-segmentation). Quantitative results of several recent methods on SIFTflow dataset for scene labeling are also shown in Table 5.2, including parametric and non-parametric methods, where non-parametric methods typically involve image retrieval process for label transfer. We evaluate the pixel accuracy (percentage of correctly classified pixels) and the class accuracy (average pixel accuracies per class) over all classes in this section.

Table 5.2: Quantitative comparison on the SIFTflow dataset.

| | Pixel (%) | Class (%) |
|---|---|---|
| Multiscale ConvNet, Farabet et al. 2013 [84] | 67.9 | 45.9 |
| [84] + cover (balanced frequencies) | 72.3 | 50.8 |
| [84] + cover (natural frequencies) | 78.5 | 29.6 |
| CNN, Pinheiro et al. 2014 [75] | 76.5 | 30.0 |
| Recurrent CNN, Pinheiro et al. 2014 [75] | 77.7 | 29.8 |
| Superparsing, Tighe et al. 2010 [88] | 76.9 | 29.4 |
| Tighe et al. 2013 [89] | 78.6 | 39.2 |
| Singh et al. 2013 [87] | 79.2 | 33.8 |
| Gatta et al. 2014 [177] | 78.7 | 32.1 |
| Gould et al. 2014 [92] | 78.4 | 25.7 |
| Yang et al. 2014 [93] | 79.8 | **48.7** |
| Integration Model, Shuai et al. 2015 [13] | 79.8 | 39.1 |
| [13] + metric learning | 80.1 | 39.7 |
| Shuai et al. 2016 [175] | 81.0 | 44.6 |
| [175] + metric learning | 81.2 | 45.5 |
| RCSL-I (random sampling) | 79.9 | 40.1 |
| RCSL-II (scene assisted retrieval) | 80.8 | 41.2 |
| RCSL-Seg | **81.6** | 47.5 |

As we can see that our baseline method [13] is able to achieve competitive performance except the class accuracy from [93]. Numerical results show effectiveness of the integration framework for scene labeling. Our proposed RCSL-II is able to achieve higher pixel and class accuracy compared to [13] and RCSL-I, which demonstrates that our scene assisted retrieval method and the rare class balanced complementary CNN structure is able to improve the baseline method regarding to unbalanced training data problems for rare class objects.

Since pre-segmentation has been used in [93] for superpixel extraction, better class accuracy may be achieved compared to [13] where only small image patches are used for learning process, and no object boundaries have been considered during the training. Consequently, with the simple superpixels-based re-segmentation (RCSL-Seg), we are able to achieve higher overall pixel accuracy compared with [93], and reasonably better per-class accuracy.

### 5.5.3 Comparative Study on Rare Classes

To further demonstrate the effectiveness of proposed methods, we summarize pixel and class accuracy in Table 5.3 only for rare classes in the SIFTflow dataset. It has been shown that with the global scene assisted retrieval, we can improve the classification accuracy compared to [13]. Instead of randomly selecting rare classes to acquire the retrieval subset, selection of rare class is constrained by scene level information, so that related rare classes image patches have a higher chance to appear while irrelevant rare classes might be filtered. Moreover, with the help of superpixels-based re-segmentation, we can make a fair comparison with [93], which further demonstrate the robustness of our proposed RCSL framework for correctly labeling and segmentation, especially for rare classes. Moreover, per-class labeling accuracies are shown in Fig. 5.9, where the RCSL-Seg achieves significant improvement for rare classes compared to the baseline method. But performance on common classes, mainly in the background (sky, mountain, etc.), slightly drops, which occurs due to the fact that the additional rare class balanced CNN may have some over-fitting problem due to more rare class training data.

Table 5.3: Quantitative comparison on rare classes of the SIFTflow dataset.

| | Pixel (%) | Class (%) |
|---|---|---|
| Tighe et al. 2013 [89] | 48.8 | 29.9 |
| Yang et al. 2014 [93] | 59.4 | 41.9 |
| Shuai et al. 2015 [13] | - | 30.7 |
| Shuai et al. 2016 [175] | - | 37.6 |
| RCSL-II | 61.3 | 39.2 |
| RCSL-Seg | **62.6** | **42.3** |

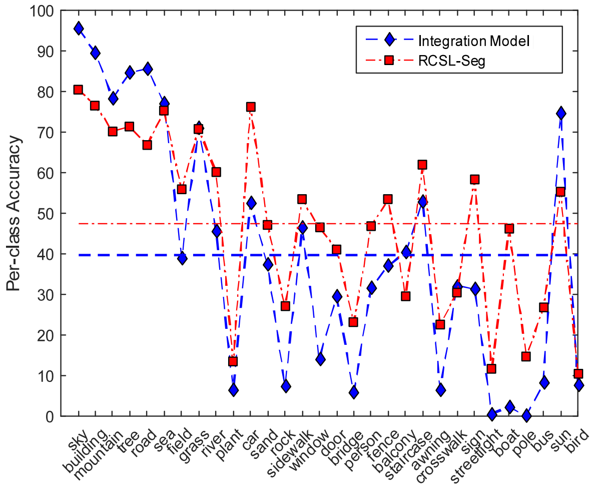


Figure 5.9: Per-class accuracy of the integration model [13] and RCSL-Seg [15].

Some qualitative examples of our proposed labeling results are shown in Figure 5.10. We can see clearly that the RCSL is more robust in finding rare classes objects, and the superpixels refinement in RCSL-Seg can further encourage the final segmentation results to be much more closer to the real world objects, and hence a higher visualizability could be reached for visual distinctions between objects. For example in the first image, we can separate the two persons instead of a rough region

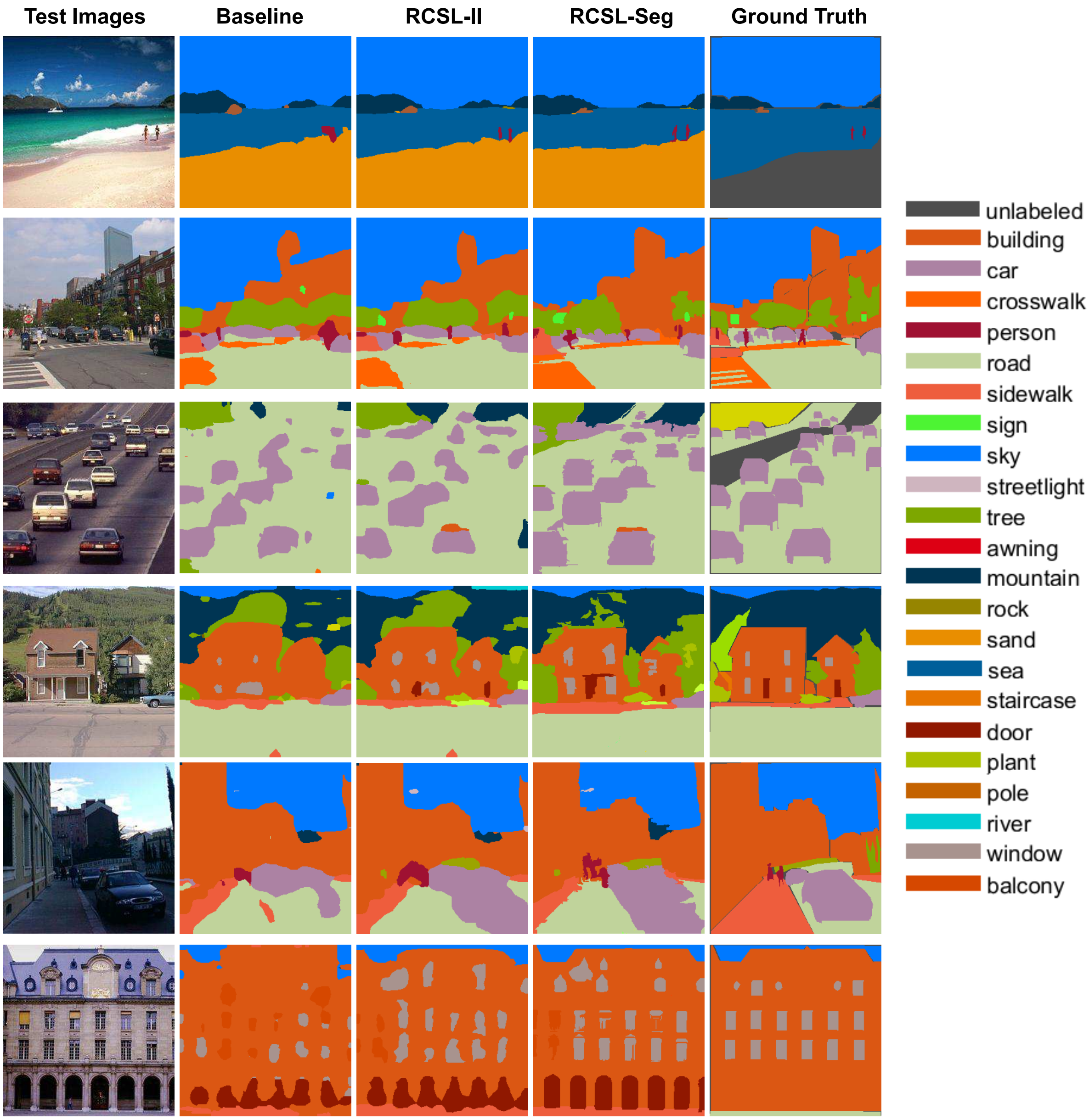


Figure 5.10: Some examples of our labeling results. The 1st column is the input images, the 2nd column is the labeling results from [13], the 3rd and the 4th columns are our proposed RCSL and RCSL-Seg labeling results respectively, while the 5th column is the ground truth labeling map.

of segmentation, and human silhouettes can be well recognized. In the third image, vehicles on the high way tend to have more detailed boundaries, and in the sixth image, we are able to recognized almost all windows of the building. Take a closer look at the second image, we are able to find the person behind the right-most car, even it is not labeled in the ground truth labeling map. Same situation happens to the streetlight appears right next to the building on the left part of the fifth image. Failed cases occasionally happen when wrong scene semantic has been discovered during the retrieval step.

## 5.6 Discussion

In this chapter, we presented a compact rare class-oriented scene labeling framework (RCSL) with global scene assisted rare classes retrieval process. Specifically we expand the retrieved image exemplar subset by choosing scene information assisted rare class image patches with less ambiguities from similar class features. And also an extra rare class balanced CNN structure is trained to mitigate imbalanced data problem, which is implemented near potential rare class pixel regions for reduced effect on frequent classes at lower computation cost. Furthermore, RCSL with superpixels-based re-segmentation (RCSL-Seg) was implemented to produce more perceptually important object boundaries. Experimental results demonstrate the effectiveness of proposed framework on both pixel and class accuracy for scene labeling tasks on the SIFTflow dataset. In the future we aim to include scene-level information rather than only CNN features to encourage more precise image retrieval, which is a very important step for the non-parametric label transfer. And moreover, since CNN is trained using image patches without considering perceptually meaningful object boundaries, further consideration of superpixels during the non-parametric labeling process will be studied.

# CHAPTER VI

# CONCLUSION AND FUTURE WORK

In this dissertation, we explore the problem of perceptually understanding a given scene by studying on object segmentation/recognition algorithms, and scene labeling methods respectively. First, we have integrated a shape generative model (CVIM) into a probabilistic level set framework to implement joint target recognition, segmentation and pose estimation in IR imagery. Due to the multi-modal nature of the optimization problem, we first implemented a PSO-based method to jointly optimize CVIM-based implicit shape matching and level set segmentation and then developed the gradient-boosted PSO (GB-PSO) algorithm to further improve the efficiency by taking advantage of the analytical and differentiable nature of CVIM. We have conducted two comparative studies on the recent SENSIAC ATR database to demonstrate the advantages of the two PSO-based algorithms. The first study involves five methods where CVIM is optimized by either explicit shape matching or MCMC-based implicit shape matching. GB-PSO and PSO are shown to be more effective than other methods. Moreover, GB-PSO was also shown to be more efficient than PSO with a much improved convergence rate due to the gradient-driven technique. The second study includes a few recent ATR algorithms for performance evaluation. It is shown that the proposed GB-PSO algorithm moderately outperforms other recent ATR algorithms at a much lower computational load. The proposed framework could be further extended to incorporate new appearance features or effective optimization to deal with more challenging ATR problems.

On the other hand, we implemented some sparse representation classification methods on a more challenging dataset, and in order to achieve higher recognition accuracy, we designed a switchable sparse representation classification method (SSRC) which fuses the traditional sparse representation classification method and a multi-attribute sparse representation classification method (MALGC) involving an sparsity concentration index regulation. Quantitative results on the Comanche dataset show promising performances of proposed fusion framework. Currently group constraints from different attributes could be incorporated into the optimization problem as a typical $l_1$ solution, but with such constraint many correctly classified testing data by the original SRC are mis-classified here though. So in the future we may explore deep learning techniques on ATR applications especially on FLIR data, to further develop an ATR system with improved efficiency and robustness.

Furthermore, we presented a compact rare classes-oriented scene labeling framework (RCSL) with a global scene assisted rare classes retrieval process, where the retrieval subset was expanded by choosing scene regulated rare class patches. A complementary rare classes focused CNN structure is trained to alleviate imbalanced data distribution problem at lower cost. moreover, a superpixels-based re-segmentation was implemented to produce more perceptually important object boundaries. Quantitative results demonstrate the promising performances of proposed framework on both pixel and class accuracy for scene labeling on the SIFTflow dataset, and improvements on rare classes labeling accuracy could be observed. In the future we are going to include not only CNN features but also scene-level information to encourage precise image retrieval, which is a very important step towards the non-parametric label transfer. Since CNN is trained using image patches without considering perceptually meaningful object boundaries, further consideration of superpixels during the non-parametric labeling process will be studied to preserve better perceptually and semantically meaning object segmentation.

## REFERENCES

[1] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Proc. CVPR*, pp. 1972–1979, June 2009.

[2] D. E. Dudgeon and R. T. Lacoss, "An overview of automatic target recognition," *Lincoln Laboratory J.*, vol. 6, no. 1, pp. 3–10, 1993.

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

[4] V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Sparsity-motivated automatic target recognition," *Appl. Opt.*, vol. 50, pp. 1425–1433, Apr. 2011.

[5] B. Rosenhahn, T. Brox, and J. Weickert, "Three-dimensional shape knowledge for joint image segmentation and pose tracking," *International Journal of Computer Vision*, vol. 73, no. 3, pp. 243–262, 2007.

[6] A. Toshev, A. Makadia, and K. Daniilidis, "Shape-based object recognition in videos using 3d synthetic object models," in *Proc. CVPR*, pp. 288–295, 2009.

[7] V. Prisacariu and I. Reid, "Nonlinear shape manifolds as shape priors in level set segmentation and tracking," in *Proc. CVPR*, pp. 2185–2192, 2011.

[8] L. J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Proc. CVPR*, pp. 2036–2043, June 2009.

[9] V. Venkataraman, G. Fan, L. Yu, X. Zhang, W. Liu, and J. Havlicek, "Automated target tracking and recognition using coupled view and identity manifolds

for shape representation," *EURASIP J. Advances in Signal Process.*, vol. 2011, no. 1, pp. 1–17, 2011.

[10] C. Bibby and I. Reid, "Robust real-time visual tracking using pixel-wise posteriors," in *Computer Vision – ECCV 2008*, pp. 831–844, Springer Berlin Heidelberg, 2008.

[11] C.-K. Chiang, T.-F. Su, C. Yen, and S.-H. Lai, "Multi-attribute sparse representation with group constraints for face recognition under different variations," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, April 2013.

[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov 1998.

[13] B. Shuai, G. Wang, Z. Zuo, B. Wang, and L. Zhao, "Integrating parametric and non-parametric models for scene labeling," in *Proc. CVPR*, pp. 4249–4258, June 2015.

[14] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[15] L. Yu and G. Fan, "Rare class oriented scene labeling using cnn incorporated label transfer," in *Advances in Visual Computing: 12th ISVC*, pp. 309–320, Springer International Publishing, 2016.

[16] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 712 – 722, 2010.

[17] A. Torralba, “Contextual priming for object detection,” *International Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[18] M. Potter, “Meaning in visual search,” *Science*, vol. 187, no. 4180, pp. 965–966, 1975.

[19] P. G. Schyns and A. Oliva, “From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition,” *Psychological Science*, vol. 5, no. 4, pp. 195–200, 1994.

[20] I. Biederman, “Perceiving real-world scenes,” *Science*, vol. 177, no. 4043, pp. 77–80, 1972.

[21] A. Oliva and A. Torralba, “The role of context in object recognition,” *Trends in cognitive sciences*, vol. 11, no. 12, pp. 520–527, 2007.

[22] B. Bhanu, “Automatic target recognition: State of the art survey,” *IEEE Trans. Aerospace and Electronic Systems*, vol. AES-22, no. 4, pp. 364–379, 1986.

[23] B. Bhanu and T. Jones, “Image understanding research for automatic target recognition,” *IEEE Aerospace, Electronic Syst. Magazine*, vol. 8, no. 10, pp. 15–23, 1993.

[24] J. A. Ratches, “Review of current aided/automatic target acquisition technology for military target acquisition tasks,” *Optical Engineering*, vol. 50, no. 7, pp. 072001–1–072001–8, 2011.

[25] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” *ACM Comput. Surv.*, vol. 38, no. 4, p. 13, 2006.

[26] U. Srinivas, V. Monga, and V. Riasati, “A comparative study of basis selection techniques for automatic target recognition,” in *Proc. IEEE Radar Conf.*, pp. 711–713, 2012.

[27] V. Bhatnagar, A. Shaw, and R. Williams, "Improved automatic target recognition using singular value decomposition," in *Proc. 1998 IEEE Int'l. Conf. Acoust., Speech, Signal Process.*, vol. 5, pp. 2717–2720, 1998.

[28] C. Olson and D. Huttenlocher, "Automatic target recognition by matching oriented edge pixels," *IEEE Trans. Image Processing*, vol. 6, no. 1, pp. 103–113, 1997.

[29] D. P. Casasent, J. S. Smokelin, and A. Ye, "Wavelet and gabor transforms for detection," *Optical Engineering*, vol. 31, no. 9, pp. 1893–1898, 1992.

[30] U. Grenander, M. Miller, and A. Srivastava, "Hilbert-schmidt lower bounds for estimators on matrix Lie groups for atr," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 790–802, 1998.

[31] L.-C. Wang, S. Z. Der, N. M. Nasrabadi, and S. A. Rizvi, "Automatic target recognition using neural networks," in *Proc. SPIE*, vol. 3466, pp. 278–289, 1998.

[32] W. Xiong and L. Cao, "Automatic target recognition based on rough set-support vector machine in SAR images," in *Proc. Int'l. Joint Conf. on Computational Sciences and Optimization*, vol. 1, pp. 489–491, 2009.

[33] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[34] L. Dong, X. Yu, L. Li, and J. K. E. Hoe, "Hog based multi-stage object detection and pose recognition for service robot," in *11th International Conference on Control Automation Robotics Vision*, pp. 2495–2500, Dec 2010.

[35] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, vol. 2, pp. 1150–1157, 1999.

[36] A. Johnson and M. Hebert, “Using spin images for efficient object recognition in cluttered 3D scenes,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999.

[37] A. Opelt, A. Pinz, and A. Zisserman, “Learning an alphabet of shape and appearance for multi-class object detection,” *International Journal of Computer Vision*, vol. 80, no. 1, pp. 16–44, 2008.

[38] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[39] Q. Mo and B. A. Draper, “Semi-nonnegative matrix factorization for motion segmentation with missing data,” in *Computer Vision – ECCV 2012*, pp. 402–415, Springer Berlin Heidelberg, 2012.

[40] M. Heikkila and M. Pietikainen, “A texture-based method for modeling the background and detecting moving objects,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–662, 2006.

[41] P. Ricaurte, C. Chilán, C. A. Aguilera-Carrasco, B. X. Vintimilla, and A. D. Sappa, “Feature point descriptors: Infrared and visible spectra,” *Sensors*, vol. 14, no. 2, pp. 3690–3701, 2014.

[42] H. V. Nguyen and F. Porikli, “Support vector shape: A classifier-based shape representation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 970–982, 2013.

[43] M. N. Saidi, A. Toumi, B. Hoeltzener, A. Khenchaf, and D. Aboutajdine, “Aircraft target recognition: A novel approach for features extraction from ISAR images,” in *International Radar Conference ”Surveillance for a Safer World” (RADAR)*, pp. 1–5, 2009.

[44] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[45] C. Xu and J. Prince, “Snakes, shapes, and gradient vector flow,” *IEEE Trans. Image Processing*, vol. 7, no. 3, pp. 359–369, 1998.

[46] N. Paragios and R. Deriche, “Geodesic active contours and level sets for the detection and tracking of moving objects,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 3, pp. 266–280, 2000.

[47] G. B. Unal and A. H. Krim, “Segmentation and target recognition in sar imagery using a level-sets-multiscale-filtering technique,” in *Proc. SPIE*, vol. 4391, pp. 370–379, 2001.

[48] M. Chen and J. Cai, “An on-line learning tracking of non-rigid target combining multiple-instance boosting and level set,” in *Proc. SPIE*, vol. 8918, 2013.

[49] M. Leventon, W. Grimson, and O. Faugeras, “Statistical shape influence in geodesic active contours,” in *Proc. CVPR*, pp. 316–323, 2000.

[50] A. Tsai, J. Yezzi, A., W. Wells, C. Tempany, D. Tucker, A. Fan, W. Grimson, and A. Willsky, “A shape-based approach to the segmentation of medical imagery using level sets,” *IEEE Trans. Medical Imaging*, vol. 22, pp. 137–154, Feb. 2003.

[51] C. Lee and A. Elgammal, “Modeling view and posture manifolds for tracking,” in *Proc. ICCV*, 2007.

[52] V. Prisacariu and I. Reid, “Shared shape spaces,” in *Proc. ICCV*, pp. 2587–2594, 2011.

[53] J. Gong, G. Fan, L. Yu, J. P. Havlicek, D. Chen, and N. Fan, "Joint target tracking, recognition and segmentation for infrared imagery using a shape manifold-based level set," *Sensors*, vol. 14, no. 6, pp. 10124–10145, 2014.

[54] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int'l. Conf. Neural Networks*, vol. 4, pp. 1942–1948, Nov. 1995.

[55] M. M. Noel, "A new gradient based particle swarm optimization algorithm for accurate computation of global minimum," *Applied Soft Computing*, vol. 12, no. 1, pp. 353–359, 2012.

[56] L. Yu, G. Fan, J. Gong, and J. P. Havlicek, "Joint infrared target recognition and segmentation using a shape manifold-aware level set," *Sensors*, vol. 15, no. 5, pp. 10118–10145, 2015.

[57] L. Yu, G. Fan, J. Gong, and J. Havlicek, "Simultaneous target recognition, segmentation and pose estimation," in *Proc. ICIP*, pp. 2655–2659, 2013.

[58] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Computer Vision – ECCV 2008*, pp. 44–57, Springer Berlin Heidelberg, 2008.

[59] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Scene parsing with multiscale feature learning, purity trees, and optimal covers," in *Proc. International Conference on Machine Learning (ICML)*, vol. 1, pp. 575–582, 2012.

[60] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Proc. CVPR*, pp. 2141–2148, June 2010.

[61] P. Kontschieder, S. R. Bulò, M. Pelillo, and H. Bischof, "Structured labels in random forests for semantic labelling and object detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, pp. 2104–2116, Oct 2014.

[62] C. Zhang, L. Wang, and R. Yang, “Semantic segmentation of urban scenes using dense depth maps,” in *Computer Vision – ECCV 2010*, pp. 708–721, Springer Berlin Heidelberg, 2010.

[63] J. Tighe, M. Niethammer, and S. Lazebnik, “Scene parsing with object instance inference using regions and per-exemplar detectors,” *International Journal of Computer Vision*, vol. 112, no. 2, pp. 150–171, 2015.

[64] A. Sharma, O. Tuzel, and M.-Y. Liu, “Recursive context propagation network for semantic scene labeling,” in *Advances in Neural Information Processing Systems 27*, pp. 2447–2455, Curran Associates, Inc., 2014.

[65] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, “Scene labeling with lstm recurrent neural networks,” in *Proc. CVPR*, pp. 3547–3555, June 2015.

[66] R. Mottaghi, S. Fidler, J. Yao, R. Urtasun, and D. Parikh, “Analyzing semantic segmentation using hybrid human-machine crfs,” in *Proc. CVPR*, pp. 3143–3150, June 2013.

[67] R. Mottaghi, X. Chen, X. Liu, N. G. Cho, S. W. Lee, S. Fidler, R. Urtasun, and A. Yuille, “The role of context for object detection and semantic segmentation in the wild,” in *Proc. CVPR*, pp. 891–898, June 2014.

[68] S. Gould, R. Fulton, and D. Koller, “Decomposing a scene into geometric and semantically consistent regions,” in *Proc. ICCV*, pp. 1–8, Sept 2009.

[69] D. Larlus and F. Jurie, “Combining appearance models and markov random fields for category level object segmentation,” in *Proc. CVPR*, pp. 1–7, June 2008.

[70] B. Triggs and J. J. Verbeek, "Scene segmentation with crfs learned from partially labeled images," in *Advances in Neural Information Processing Systems 20*, pp. 1553–1560, Curran Associates, Inc., 2008.

[71] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Computer Vision – ECCV 2006*, pp. 1–15, Springer Berlin Heidelberg, 2006.

[72] Y. Zhang and T. Chen, "Efficient inference for fully-connected crfs with stationarity," in *Proc. CVPR*, pp. 582–589, June 2012.

[73] X. He, R. S. Zemel, and M. A. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Proc. CVPR*, vol. 2, pp. II–695–II–702 Vol.2, June 2004.

[74] L. Ladický, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical crfs for object class image segmentation," in *Proc. ICCV*, pp. 739–746, Sept 2009.

[75] P. H. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. International Conference on Machine Learning (ICML)*, pp. 82–90, 2014.

[76] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, pp. 541–551, Dec. 1989.

[77] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, Curran Associates, Inc., 2012.

[78] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, pp. 818–833, Springer International Publishing, 2014.

[79] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - weakly-supervised learning with convolutional neural networks," in *Proc. CVPR*, pp. 685–694, June 2015.

[80] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 922–928, Sept 2015.

[81] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proc. CVPR*, pp. 3982–3991, June 2015.

[82] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *Proc. ICCV*, pp. 1134–1142, Dec 2015.

[83] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proc. CVPR*, pp. 2155–2162, June 2014.

[84] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1915–1929, Aug 2013.

[85] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2368–2382, Dec 2011.

[86] D. Eigen and R. Fergus, "Nonparametric image parsing using adaptive neighbor sets," in *Proc. CVPR*, pp. 2799–2806, June 2012.

[87] G. Singh and J. Kosecka, "Nonparametric scene parsing with adaptive feature relevance and semantic context," in *Proc. CVPR*, pp. 3151–3157, June 2013.

[88] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *Computer Vision – ECCV 2010*, pp. 352–365, Springer Berlin Heidelberg, 2010.

[89] J. Tighe and S. Lazebnik, "Finding things: Image parsing with regions and per-exemplar detectors," in *Proc. CVPR*, pp. 3001–3008, June 2013.

[90] F. Tung and J. J. Little, "Collageparsing: Nonparametric scene parsing by adaptive overlapping windows," in *Computer Vision – ECCV 2014*, pp. 511–525, Springer International Publishing, 2014.

[91] S. Gould and Y. Zhang, "Patchmatchgraph: Building a graph of dense patch correspondences for label transfer," in *Computer Vision – ECCV 2012*, pp. 439–452, Springer Berlin Heidelberg, 2012.

[92] S. Gould, J. Zhao, X. He, and Y. Zhang, "Superpixel graph label transfer with learned distance metric," in *Computer Vision – ECCV 2014*, pp. 632–647, Springer International Publishing, 2014.

[93] J. Yang, B. Price, S. Cohen, and M. H. Yang, "Context driven scene parsing with attention to rare classes," in *Proc. CVPR*, pp. 3294–3301, June 2014.

[94] G. Heitz, S. Gould, A. Saxena, and D. Koller, "Cascaded classification models: Combining models for holistic scene understanding," in *Advances in Neural Information Processing Systems 21*, pp. 641–648, Curran Associates, Inc., 2009.

[95] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr, "Combining appearance and structure from motion features for road scene understanding," in *Proc. British Machine Vision Conference*, pp. 62.1–62.11, BMVA Press, 2009.

[96] M. George, “Image parsing with a wide range of classes and scene-level context,” in *Proc. CVPR*, pp. 3622–3630, June 2015.

[97] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, pp. 359–366, Jul. 1989.

[98] V. Mirelli and S. A. Rizvi, “Automatic target recognition using a multilayer convolution neural network,” in *Proc. SPIE*, vol. 2755, pp. 106–125, 1996.

[99] L. Wang, S. Der, and N. Nasrabadi, “Automatic target recognition using a feature-decomposition and data-decomposition modular neural network,” *IEEE Trans. Image Processing*, vol. 7, no. 8, pp. 1113–1121, 1998.

[100] L. Chan, N. Nasrabadi, and V. Mirelli, “Multi-stage target recognition using modular vector quantizers and multilayer perceptrons,” in *Proc. CVPR*, pp. 114–119, 1996.

[101] L. A. Chan and N. M. Nasrabadi, “Automatic target recognition using vector quantization and neural networks,” *Optical Engineering*, vol. 38, no. 12, pp. 2147–2161, 1999.

[102] P. Christiansen, K. A. Steen, R. N. Jørgensen, and H. Karstoft, “Automated detection and recognition of wildlife using thermal cameras,” *Sensors*, vol. 14, no. 8, pp. 13778–13793, 2014.

[103] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[104] Z.-Z. Li, J. Chen, Q. Hou, H.-X. Fu, Z. Dai, G. Jin, R.-Z. Li, and C.-J. Liu, “Sparse representation for infrared dim target detection via a discriminative

over-complete dictionary learned online," *Sensors*, vol. 14, no. 6, pp. 9451–9470, 2014.

[105] X. Li, R. Guo, and C. Chen, "Robust pedestrian tracking and recognition from FLIR video: A unified approach via sparse coding," *Sensors*, vol. 14, no. 6, pp. 11245–11259, 2014.

[106] A. Srivastava, "Bayesian filtering for tracking pose and location of rigid targets," in *Proc. SPIE*, vol. 4052, pp. 160–171, 2000.

[107] V. Prisacariu and I. Reid, "Pwp3d: Real-time segmentation and tracking of 3d objects," in *Proc. British Machine Vision Conference*, pp. 47.1–47.10, BMVA Press, 2009.

[108] F. A. Sadjadi, "Automatic object recognition: critical issues and current approaches," in *Proc. SPIE*, vol. 1471, pp. 303–313, 1991.

[109] A. Aull, R. Gabel, and T. Goblick, "Real-time radar image understanding: A machine intelligence approach," *Lincoln Laboratory J.*, vol. 5, no. 2, pp. 195–222, 1992.

[110] J. Liebelt, C. Schmid, and K. Schertler, "Viewpoint-independent object class detection using 3d feature maps," in *Proc. CVPR*, 2008.

[111] S. Khan, H. Cheng, D. Matthies, and H. Sawhney, "3d model based vehicle classification in aerial imagery," in *proc. CVPR*, pp. 1681–1687, 2010.

[112] G. Paravati and S. Esposito, "Relevance-based template matching for tracking targets in FLIR imagery," *Sensors*, vol. 14, no. 8, pp. 14106–14130, 2014.

[113] S. K. Chari, C. E. Halford, and E. Jacobs, "Human target identification and automated shape based target recognition algorithms using target silhouette," in *Proc. SPIE*, vol. 6941, pp. 69410B–1–69410B–9, 2008.

[114] J. Xiao and M. Shah, "Automatic target recognition using multi-view morphing," in *Proc. SPIE*, vol. 5426, pp. 391–399, 2004.

[115] M. Chahooki and N. Charkari, "Learning the shape manifold to improve object recognition," *Machine Vision and Applications*, vol. 24, no. 1, pp. 33–46, 2013.

[116] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. CVPR*, pp. 3485–3492, June 2010.

[117] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1075–1088, Sept 2003.

[118] M. Liang, X. Hu, and B. Zhang, "Convolutional neural networks with intra-layer recurrent connections for scene labeling," in *Advances in Neural Information Processing Systems 28*, pp. 937–945, Curran Associates, Inc., 2015.

[119] S. Bulò and P. Kontschieder, "Neural decision forests for semantic image labelling," in *Proc. CVPR*, pp. 81–88, June 2014.

[120] P. Kohli, L. Ladicky, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," in *Proc. CVPR*, pp. 1–8, June 2008.

[121] Q. Huang, M. Han, B. Wu, and S. Ioffe, "A hierarchical conditional random field model for labeling and segmenting images of street scenes," in *Proc. CVPR*, pp. 1953–1960, June 2011.

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

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

[124] T. Jebara, "Images as bags of pixels," in *proc. ICCV*, pp. 265–272, 2003.

[125] D. Cremers, "Dynamical statistical shape priors for level set-based tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1262–1273, Aug. 2006.

[126] S. Das, A. Abraham, and A. Konar, "Particle swarm optimization and differential evolution algorithms: Technical analysis, applications and hybridization perspectives," in *Advances of Computational Intelligence in Industrial Systems*, vol. 116 of *Studies in Computational Intelligence*, pp. 1–38, Springer Berlin Heidelberg, 2008.

[127] J. Liang, A. Qin, P. Suganthan, and S. Baskar, "Comprehensive learning particle swarm optimizer for global optimization of multimodal functions," *IEEE Trans. Evolutionary Computation*, vol. 10, pp. 281–295, June 2006.

[128] S. Esquivel and C. A. Coello Coello, "On the use of particle swarm optimization with multimodal functions," in *The 2003 Congress on Evolutionary Computation*, vol. 2, pp. 1130–1136 Vol.2, Dec 2003.

[129] R. Poli, "Analysis of the publications on the applications of particle swarm optimisation," *J. Artif. Evol. App.*, vol. 2008, pp. 4:1–4:10, Jan. 2008.

[130] M. Benedetti, R. Azaro, D. Franceschini, and A. Massa, "Pso-based real-time control of planar uniform circular arrays," *IEEE Antennas and Wireless Propagation Letters*, vol. 5, pp. 545–548, Dec 2006.

[131] H. Elkamchouchi and M. Wagih, "Dynamic null steering in linear antenna arrays using adaptive particle swarm optimization algorithm," in *International Conference on Wireless and Mobile Communications*, pp. 24–24, March 2007.

[132] D. Gies and Y. Rahmat-Samii, "Reconfigurable array design using parallel particle swarm optimization," in *IEEE Antennas and Propagation Society International Symposium*, vol. 1, pp. 177–180 vol.1, June 2003.

[133] R. Xu and D. Wunsch, "Gene regulatory networks inference with recurrent neural network models," in *IEEE International Joint Conference on Neural Networks*, vol. 1, pp. 286–291 vol. 1, July 2005.

[134] R. Xu, D. Wunsch, and R. Frank, "Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 4, pp. 681–692, Oct 2007.

[135] A. Babazadeh, H. Poorzahedy, and S. Nikoosokhan, "Application of particle swarm optimization to transportation network design problem," *Journal of King Saud University - Science*, vol. 23, no. 3, pp. 293 – 300, 2011. Special Issue on Advances in Transportation Science.

[136] L. Yan and J. Zeng, "Using particle swarm optimization and genetic programming to evolve classification rules," in *6th World Congress on Intelligent Control and Automation*, vol. 1, pp. 3415–3419, 2006.

[137] E. Miguelanez, A. Zalzala, and P. Buxton, "Swarm intelligence in automated electrical wafer sort classification," in *IEEE Congress on Evolutionary Computation*, vol. 2, pp. 1597–1604 Vol. 2, Sept 2005.

[138] P. de Moura Oliveira, “Modern heuristics review for pid control optimization: a teaching experiment,” in *International Conference on Control and Automation*, vol. 2, pp. 828–833 Vol. 2, June 2005.

[139] Z.-L. Gaing, “A particle swarm optimization approach for optimum design of pid controller in avr system,” *IEEE Trans. Energy Conversion*, vol. 19, pp. 384–391, June 2004.

[140] J. Nenortaite and R. Simutis, “Adapting particle swarm optimization to stock markets,” in *Proc. International Conference on Intelligent Systems Design and Applications*, pp. 520–525, Sept 2005.

[141] Y. Song, Z. Chen, and Z. Yuan, “New chaotic pso-based neural network predictive control for nonlinear process,” *IEEE Trans. Neural Networks*, vol. 18, pp. 595–601, March 2007.

[142] C.-M. Huang and F.-L. Wang, “An rbf network with ols and epso algorithms for real-time power dispatch,” *IEEE Trans. Power Systems*, vol. 22, pp. 96–104, Feb 2007.

[143] C.-F. Juang, “A hybrid of genetic algorithm and particle swarm optimization for recurrent network design,” *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, pp. 997–1006, April 2004.

[144] J. Vlachogiannis and K. Lee, “A comparative study on particle swarm optimization for optimal steady-state performance of power systems,” *IEEE Trans. Power Systems*, vol. 21, pp. 1718–1728, Nov 2006.

[145] A. Chatterjee, K. Pulasinghe, K. Watanabe, and K. Izumi, “A particle-swarm-optimized fuzzy-neural network for voice-controlled robot systems,” *IEEE Trans. Industrial Electronics*, vol. 52, pp. 1478–1489, Dec 2005.

[146] G. Venayagamoorthy and W. Zha, “Comparison of nonuniform optimal quantizer designs for speech coding with adaptive critics and particle swarm,” *IEEE Trans. Industry Applications*, vol. 43, pp. 238–244, Jan 2007.

[147] H.-G. Sun, Y.-X. Pan, and Y.-F. Zhang, “Apso based gabor wavelet feature extraction method,” in *Proc. International Conference on Machine Learning and Cybernetics*, vol. 6, pp. 3888–3893 vol.6, Aug 2004.

[148] G. Evers and M. Ben Ghalia, “Regrouping particle swarm optimization: A new global optimization algorithm with improved performance consistency across benchmarks,” in *Proc. IEEE Int’l. Conf. Systems, Man and Cybernetics*, pp. 3901–3908, Oct. 2009.

[149] M. O. Vasilescu and D. Terzopoulos, “Multilinear analysis of image ensembles: Tensorfaces,” in *Computer Vision – ECCV 2002*, pp. 447–460, Springer Berlin Heidelberg, 2002.

[150] M. Raydan and B. F. Svaiter, “Relaxed steepest descent and Cauchy-Barzilai-Borwein method,” *Computational Optimization and Applications*, vol. 21, no. 2, pp. 155–167, 2002.

[151] Z. Zivkovic and F. van der Heijden, “Efficient adaptive density estimation per image pixel for the task of background subtraction,” *Pattern Recog. Lett.*, vol. 27, pp. 773–780, 2006.

[152] E. Peli, “Contrast in complex images,” *Journal of the Optical Society of America*, vol. 7, pp. 2032–2040, 1990.

[153] M. Varma and A. Zisserman, “A statistical approach to material classification using image patch exemplars,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, pp. 2032–2047, Nov. 2009.

[154] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, pp. 898–916, May 2011.

[155] K. O. Babalola, B. Patenaude, P. Aljabar, J. Schnabel, D. Kennedy, W. Crum, S. Smith, T. F. Cootes, M. Jenkinson, and D. Rueckert, “Comparison and evaluation of segmentation techniques for subcortical structures in brain mri,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, pp. 409–416, Springer Berlin Heidelberg, 2008.

[156] F. Ge, S. Wang, and T. Liu, “Image-segmentation evaluation from the perspective of salient object extraction,” in *Proc. CVPR*, vol. 1, pp. 1146–1153, Jun. 2006.

[157] J. Carreira and C. Sminchisescu, “Cpmc: Automatic object segmentation using constrained parametric min-cuts,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1312–1328, Jul. 2012.

[158] J. Gong, G. Fan, L. Yu, J. P. Havlicek, D. Chen, and N. Fan, “Joint view-identity manifold for infrared target tracking and recognition,” *Computer Vision and Image Understanding*, vol. 118, pp. 211–224, Jan. 2014.

[159] Y. Bengio, “Learning deep architectures for ai,” *Found. Trends Mach. Learn.*, vol. 2, pp. 1–127, Jan. 2009.

[160] Y. Bengio, A. C. Courville, and P. Vincent, “Unsupervised feature learning and deep learning: A review and new perspectives,” *CoRR*, vol. abs/1206.5538, 2012.

[161] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, Aug 2013.

[162] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep machine learning - a new frontier in artificial intelligence research [research frontier]," *IEEE Computational Intelligence Magazine*, vol. 5, pp. 13–18, Nov 2010.

[163] T. S. Lee and D. Mumford, "Hierarchical bayesian inference in the visual cortex," *J. Opt. Soc. Am. A*, vol. 20, pp. 1434–1448, Jul 2003.

[164] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards ai," *Large-scale kernel machines*, vol. 34, pp. 1–41, 2007.

[165] N. Le Roux and Y. Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural Computation*, vol. 20, pp. 1631–1649, June 2008.

[166] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[167] G. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural Networks: Tricks of the Trade*, vol. 7700 of *Lecture Notes in Computer Science*, pp. 599–619, Springer Berlin Heidelberg, 2012.

[168] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?," in *Proceedings of AISTATS 2010*, vol. 9, pp. 201–208, May 2010.

[169] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proc. CVPR*, vol. 2, pp. II–97–104 Vol.2, June 2004.

[170] J. Ngiam, Z. Chen, D. Chia, P. W. Koh, Q. V. Le, and A. Y. Ng, "Tiled convolutional neural networks," in *Advances in Neural Information Processing Systems 23*, pp. 1279–1287, Curran Associates, Inc., 2010.

[171] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *Proc. ICCV*, pp. 2146–2153, Sept 2009.

[172] "Ufldl tutorial." http://ufldl.stanford.edu/tutorial/.

[173] D. George, *How the Brain Might Work: A Hierarchical and Temporal Model for Learning and Recognition.* PhD thesis, Stanford, CA, USA, 2008.

[174] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[175] B. Shuai, Z. Zuo, G. Wang, and B. Wang, "Scene parsing with integration of parametric and non-parametric models," *IEEE Trans. Image Processing*, vol. 25, pp. 2379–2391, May 2016.

[176] A. Vedaldi and K. Lenc, "Matconvnet - convolutional neural networks for matlab," in *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.

[177] C. Gatta, A. Romero, and J. v. d. Veijer, "Unrolling loopy top-down semantic feedback in convolutional deep networks," in *Proc. CVPR Workshops*, pp. 504–511, June 2014.

# VITA

Liangjiang Yu

Candidate for the Degree of

Doctor of Philosophy

Dissertation: HYBRID MACHINE LEARNING APPROACHES FOR SCENE UNDERSTANDING: FROM SEGMENTATION AND RECOGNITION TO IMAGE PARSING

Major Field: Electrical Engineering

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Electrical Engineering at Oklahoma State University, Stillwater, Oklahoma in May, 2017.

Completed the requirements for the Master of Science in Electrical Engineering at Oklahoma State University, Stillwater, Oklahoma in 2011.

Completed the requirements for the Bachelor of Science in Communication Engineering at North University of China, Taiyuan, China in 2008.

Experience:

Graduate Research Assistant in Visual Computing and Image Processing Lab (VCIPL), School of Electrical and Computer Engineering, Oklahoma State University, August 2009 - December 2016.