Automatic Target Recognition in Infrared Imagery Using Dense HOG Features and Relevance Grouping of Vocabulary

M.N.A. Khan[†], Guoliang Fan[‡,*], Douglas R. Heisterkamp[†], Liangjiang Yu[‡]
[†] Department of Computer Science, Oklahoma State University
[‡] School of Electrical and Computer Engineering, Oklahoma State University
{mohk,guoliang.fan,douglas.r.heisterkamp,liangjiang.yu}@okstate.edu

Abstract

We study automatic target recognition (ATR) in infrared (IR) imagery by applying two recent computer vision techniques, Histogram of Oriented Gradients (HOG) and Bag-of-Words (BoW). We propose the idea of dense HOG features which are extracted from a set of high-overlapped local patches in a small IR chip and we apply a vocabulary tree that is learned from a set of training images to support efficient and scalable BoW-based ATR. We develop a relevance grouping of vocabulary (RGV) technique to improve the ATR performance by additional voting from grouped visual words. Different from traditional word grouping techniques, RGV groups visual words of the same dominant class to enhance the voting confidence in BoW-based classification. The proposed ATR algorithm is evaluated against recent sparse representation-based classification (SRC) approaches that reportedly outperform traditional methods. Experimental results on the COMANCHE IR dataset demonstrate the advantages of the newly proposed algorithm (BoW-RGV) over the recent SRC approaches.

1. Introduction

Automatic target recognition (ATR) is useful and important in many civilian and military applications. This task is often necessary when large amounts of sensor data need to be processed in a timely manner. The area under observation is often populated with natural and man-made distractions and is relatively large compared with the actual target size. For example, images captured by forward-looking infrared (FLIR) sensors, are highly influenced by atmosphere and weather conditions as well as various background clutters. In general, an ATR system can be divided into four parts [5]: (1) detection, (2) clutter rejection, (3) feature extraction and (4) classification. We will focus on the last two in this work and we are interested in applying and advancing the recent computer vision techniques to address this long-standing problem.

There have been many ATR algorithms proposed during the last few decades [4, 5]. Broadly speaking, they can be categorized as *learning-based* [7, 16] and *model-based* [6, 10] approaches. In [11], a comparative study has been conducted via experiments for some selected algorithms. The evaluation results suggest that no single approach was quite satisfactory in terms of the recognition accuracy, demanding further investigation in this area. Recently, Patel *et al*. [14] proposed a sparse representation-based classification algorithm (SRC) for infrared ATR where an l_1 minimization problem has been solved to represent a target image as a linear combination of training samples. This sparse solution reveals the class (i.e., the target type) for a given IR image chip. It was reported in [14] that the SRC-based algorithm outperforms most recent ones and will be used as the major competing method in our research.

Our research in this work is motivated by recent computer vision advancements in object recognition, image classification and image retrieval. First, we develop a BoW-based (Bag-of-Words) representation for an IR chip which involves HOG features extracted from local patches. The BoW-based representation is expected to be robust to occlusion and background clutter, thanks to its localized nature. Second, we allow local patches to be highly overlapped to capture dense HOG features. On the one hand, dense HOG features support a complete characterization of an IR chip. On the other hand, it leads to a large number of features of great redundance. Third, we use the vocabulary tree [13] to learn a set of visual words from a number of training chips, where all visual words are treated independently. Inspired by [12] where a text clustering approach was proposed to group similar words for building a thesaurus automatically, we develop a new relevance grouping of vocabulary (RGV) technique to improve BoW-based classification by additional voting from grouped visual words. In RGV, visual words are grouped according to their relevance to the

*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.

same dominant class information. In other words, several visual words in an unknown target image sharing the same dominant class information will be allowed to make an additional vote to enhance the confidence of that dominant class. Our relevance-based grouping criteria are very different from those similarity-based one in [2, 15]. In this work, we use the COMANCHE IR dataset provided by the US Army Research Lab for algorithm evaluation [14], which involve ten target types viewed under 72 different viewing angles (also known as pose angle). We will compare our proposed methods with several SRC variants and experimental results demonstrate the advantages of our algorithms in terms of both accuracy and efficiency.

2. Proposed method

In this section, we will first introduce the idea of dense HOG feature extraction. Then we present a new RGV technique to boost BoW-based classification. We also provide the pseudo code of the proposed ATR algorithm that involves a two-phase voting.

2.1. Dense HoG Feature Extraction

Recently, several local feature descriptors (SIFT, HOG, SURF, PCA-SIFT) have been proposed in the literature for a variety of computer vision applications. Especially, the HOG (Histogram of Gradient) descriptor [8] that encapsulates the gradient information of an image in orderless histograms will be used in this work due to its simplicity and robustness to noise, occlusion and variations of views and illumination. HOG features are extracted from highly-overlapped local patches which are centered on a set of densely sampled pixels in the target region, as shown in Figure 1. In our work, sampling density is fixed to 1 pixel and we refer this approach as dense HOG feature extraction.

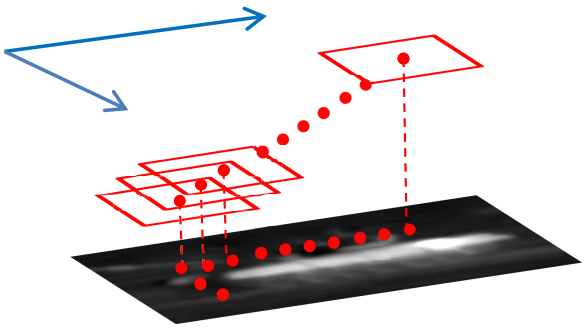

Figure 1: Dense HOG feature extraction from a set of highly overlapped local patches which are centered on densely sampled pixels (red).

2.2. RGV for BoW-based Classification

Due to the highly-overlapped nature of local patches, there is significant redundance among dense HOG features which can be quantized in a feature space to form a collection of visual words. Thus an image can be represented by a BoW that encodes the occurrences of different visual words. Each word contains some attributes pertaining to certain target type. To support scalable and fast ATR, we can construct a vocabulary tree via hierarchical K-means clustering of all dense HOG features extracted from training chips [13], where a weighting scheme (*e.g.* term frequency-inverse document frequency (*tf-idf*) weighting) is applied to the vocabulary to reduce the effect of frequently occurring words. During learning, each cluster is recursively partitioned up to a certain level which forms a hierarchical tree-like structure called the vocabulary tree. Each leaf of the tree or lowest level cluster center (centroid) represents a visual word, as shown in Figure 2.

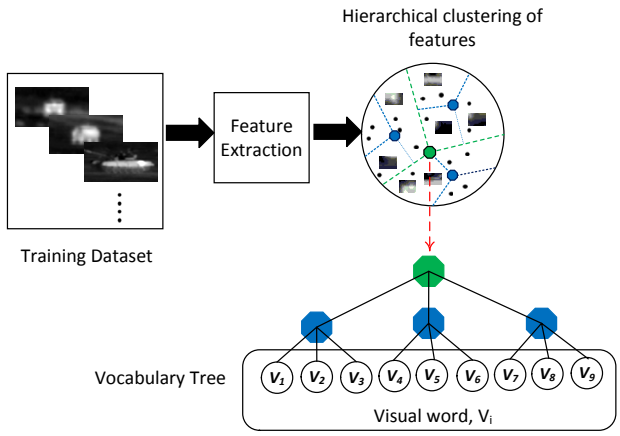

Figure 2: Steps of visual vocabulary creation. After feature extraction, hierarchical k-means clustering is applied to feature vectors. Each level of hierarchy represents an equivalent level (L) of tree structure and K represents branching factor. In this tree, $K = 3$ and $L = 3$.

Usually, BoW-based classification treats each visual word independently. Inspired by text retrieval in [12], we want to improve classification performance by exploiting the occurrence of multiple visual words sharing the same dominant class information. To do so, we need to find the dominant information for each visual word and then group the visual words of similar relevance in terms of label prediction (i.e., the target identity and pose angle in the context of IR ATR). We refer to this process as *Relevance Grouping of Vocabulary* (RGV). Finding an appropriate grouping results in an optimization problem where different criteria may be involved [2, 15]. In our work, the RGV process is performed in two sequential steps as follows:

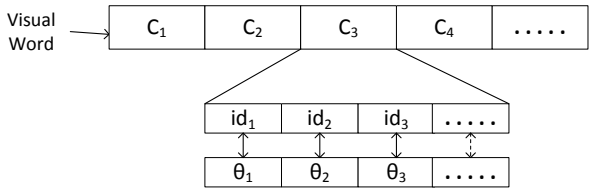

Figure 3: The layout of a visual word (c_i = class label, id_i = image ID, θ_i = pose of corresponding target in image id_i).

2.2.1 Step 1: Finding *dominant* class, T^*

As an IR image chip represents a single target class along with a specific pose angle, a visual word may be viewed as a collection of similar local features originating from images of different target classes. Therefore, each word is also associated with a number of pose angles under each of these classes (See Figure 3). Before visual word grouping, we first tag each word with the *dominant* class label. The *dominant* class of a visual word is defined as the class having the word's maximum influence. Having the same *dominant* class is a prerequisite to form a group of visual words. We use two popular metrics (*Information Gain* and *Chi-Square*) adapted from *Information Retrieval* literature to determine the *dominant* class for each visual word. For a visual word v, whose dominant class is denoted by T^*, we can find a pose distribution under the class label T^* which is also used for grouping.

Information Gain (I.G.): Assume C is a set of N classes, $C = \{c_1, c_2, \ldots, c_N\}$ and $\mathbf{V} = \{v_j\}_{j=1}^{M}$ is the visual vocabulary. According to information theory literature, I.G. is known as a measure of knowledge gained about class c_i because of the fact that word v_j is present. It considers both occurrence and non-occurrence of words in images. In other words, if c_i and v_j are independent, then information gain I.G. will be zero [3]. The information gain $I.G.^i$ of a word v_j over c_i is defined as follows [3],

$$I.G.^i(v_j) = H(c_i) - H(c_i|v_j) - H(c_i|\neg v_j). \quad (1)$$

where $H(c_i)$ is the entropy of c_i and $H(c_i|v_j)$ and $H(c_i|\neg v_j)$ are conditional entropies of c_i in the presence and absence of word v_j respectively. Assume IR chip I belongs to class c_i and the set of words representing I is denoted as I_v, we have:

$$\begin{aligned} H(c_i) &= -P(c_i)\log P(c_i), & (2)\\ H(c_i|v_j) &= -P(v_j, c_i)\log P(c_i|v_j), & (3)\\ H(c_i|\neg v_j) &= -P(\neg v_j, c_i)\log P(c_i|\neg v_j), & (4) \end{aligned}$$

where,

$$\begin{aligned} P(c_i) &= \text{Prob}(I \in c_i),\\ P(v_j, c_i) &= \text{Prob}(v_j \in I_v \text{ and } I \in c_i),\\ P(c_i|v_j) &= \text{Prob}(I \in c_i \text{ given } v_j \in I_v). \end{aligned}$$

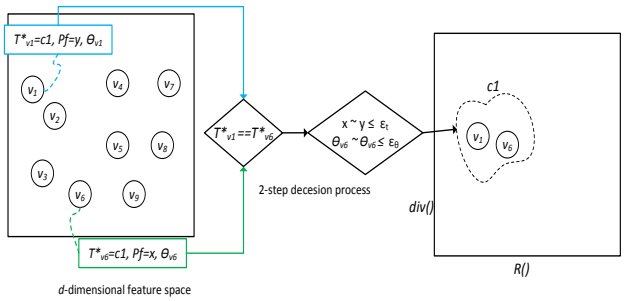

Figure 4: An illustration of RGV where each circle represents a visual word (left box) and the dotted boundary indicates the formation of a relevance group (right box).

Using the above equations, for each visual word v_j, a sequence of $I.G.$'s is computed for N classes.

Chi-Square (C.S.): It also measures dependency between word v_j and class c_i. In other words, if v_j and c_i are dependent, then the occurrence (or non-occurrence) of v_j makes the occurrence of c_i more likely (or less likely). Following the above notations, the Chi-Square metric for v_j with respect to class c_i is defined as follows [3],

$$C.S.^i(v_j) = \frac{m(P(v_j,c_i)P(\neg v_j,\neg c_i) - P(v_j,\neg c_i)P(\neg v_j,c_i))^2}{P(v_j)P(\neg v_j)P(c_i)P(\neg c_i)} \quad (5)$$

where m is the size of dataset (here, total number of IR chips). Like the previous metric, a collection of C.S.'s for word v_j can be computed for N classes.

Using any of the two metrics defined above, we can find the *dominant* class T^* of a visual word as follows,

$$T^* = \arg\max_i \{I.G.^i\}_{i=1}^{N} \text{ or } \arg\max_i \{C.S.^i\}_{i=1}^{N}.$$

This *dominant* class signifies its maximum knowledge gain towards that class over other class labels.

2.2.2 Step 2: Grouping of Visual Words

If two words u and v share the same *dominant* class, they will be considered as the *candidate* words to form a relevance group. Assume $\Omega = \{v_1, v_2, \ldots, v_k\}$ is a relevance group of visual words sharing the same *dominant* class c_p. Given two words (v_i, v_j) in Ω, they will satisfy the following conditions regarding the similarity of their knowledge gain towards the same *dominant* class and the divergence of their pose distributions (Θ_{v_i} and Θ_{v_j}):

$$\begin{aligned} |R(v_i) - R(v_j)| &\leq \epsilon_t, & (6)\\ div(\Theta_{v_i}, \Theta_{v_j}) &\leq \epsilon_\theta & (7) \end{aligned}$$

where $(v_i, v_j) \in \Omega$, $div(\cdot)$ is a divergence measure between two distributions, $R(v)$ is the I.G.-based or C.S.-based *relevance function* of a word v defined as in (1) and (5) under the *dominant* class of v, ϵ_t and ϵ_θ are thresholds chosen to ensure non-overlap grouping. In our paper, we used

Kullback-Leibler (KL) divergence [9] as *div(.)*. The whole RGV process is illustrated in Figure 4.

2.3. ATR via 2-phase Voting

The classification is accomplished by a 2-phase voting. The final ATR decision is made by combining the results from the two phases. Phase-I voting is similar to the baseline BoW approach described in Algorithm 1. In Phase-II, we will utilize the RGV groups to boost up the contribution of coexistent visual words belonging to same group.

Algorithm 1 Phase-I voting using the original BoW.

Input: Vocabulary tree, T, query image, I_q.
Output: The first N-class histogram, $S_1(c)$ with $c = (1, ..., N)$
Initialization: extract HOG descriptors from I_q using dense sampling and initialize the N-class histogram.
begin
for all features in the query image **do**
  $current\_node \leftarrow root$
  while $current\_node$ has no children **do**
    choose the best match child node v.
    $current\_node \leftarrow v$
  end while
  retrieve the associated class labels and their respective weights from the current node.
  add the weights to the bins of corresponding class labels.
end for
return the first N-class histogram.
end

Given a query image, I_q, Phase-I voting returns an N-class histogram $S_1(c)$ with $c = (1, ..., N)$ by using visual words of I_q, showing the confidence of each class towards the estimation of the target class. Phase-II voting described in Algorithm 2 provides an additional N-class histogram, $S_2(c)$ with $c = (1, ..., N)$, indicating the additional contribution from relevance groups, Π. After the 2-phase voting, we combine two N-class histograms as,

$$S(c) = S_1(c) + S_2(c), \forall c \in \{1, 2, \ldots, N\}. \quad (8)$$

The final classification (ATR) result is obtained by:

$$C_q = \arg\max_i \{S(c)\}_{i=1}^{N} \quad (9)$$

where C_q is the estimated target class for query image I_q.

3. Experiments

In this section, we evaluate the proposed RGV algorithm by comparing it with the baseline BoW method and several SRC variants that are implemented from the open-source

Algorithm 2 Phase-II voting using the RGV.

Input: Relevance groups, $\Pi = \{\Omega_1, \Omega_2, \ldots, \Omega_R\}$, and query image, I_q.
Output: The second N-class histogram, $S_2(c)$ with $c = (1, ..., N)$.
Initialization: Revisit all visual words in I_q and initialize the second N-class histogram.
begin
for all visual words in I_q belonging to Π **do**
  find their *dominant* class labels and weights.
  add the weights to the bins of corresponding class labels.
end for
return the second N-class histogram.
end

SRC package (spectral projected gradient (SPGL1) [1]). Specifically, we chose three sparse solutions, i.e., Lasso, Basis Pursuit Denoise (BPDN) and Basis Pursuit(BP) as those used in [14].

3.1. Infrared Dataset

The original Comanche IR dataset has two groups of IR chips along with ground-truth information (target type, pose *etc*.). The **SIG** group contains 13860 chips captured under relatively favorable atmospheric conditions. The dimension of each chip is 40×75 pixels. There are 10 types of military targets denoted as TG1, TG2, TG3, ..., TG10 (see Figure 5) in **SIG** and each target is viewed at 72 different orientations in azimuth (0°, 5°, $\ldots$, 360°) from a fixed range. Another group named **ROI** contains around 3300 chips consisting of five targets TG1, TG2, TG3, TG4 and TG7 where aspect angles are coarsely sampled (at 45°) under less favorable conditions. In our experiments, we only have the **SIG** dataset available. The SIG set is randomly partitioned into two non-overlapped sub-groups, **SIG-TRAIN** (10872 chips) and **SIG-TEST** (2988 chips) used for training and testing respectively. For the sake of fair performance comparison, we have used the same settings for all experiments, unless otherwise stated.

3.2. Recognition performance

The dimension of HOG feature is controlled by the number of cells and the histogram bin size. In all our experiments, the cell size is 5×7 pixels and the number of histogram bin is 5. In each image chip, a local 4×4-cell patch is defined for every pixel in the central region (14×18 pixels) of the chip and thus there are 252 dense HOG features extracted for each chip. Although there is a tremendous redundancy in this dense HOG features extraction, the hierarchical vocabulary tree can efficiently quantize them into a more compact set of visual words.

Table 1: Target recognition performance(%) using proposed RGV and baseline BoW methods at different vocabulary sizes.

Vocabulary Tree Parameters	Vocabulary Size $\|V\|$	Avg #Features/Word (Quantization ratio)	BoW	RGV with I.G.		RGV with C.S.	
				Accuracy	$\|\Pi\|$	Accuracy	$\|\Pi\|$
K=5 L=6	15,625	175.34	84.60	85.61	10256	86.08	10857
K=6 L=6	46,637	58.75	91.30	92.60	1827	93.17	8552
K=7 L=6	117,347	23.35	94.91	96.05	1271	96.39	4702
K=7 L=7	747,424	3.67	97.82	99.10	356	99.10	880

* K=Branch factor, L=#Levels, $\|\Pi\|$=the number of relevance groups, I.G.=Information Gain, C.S.=Chi-Square.

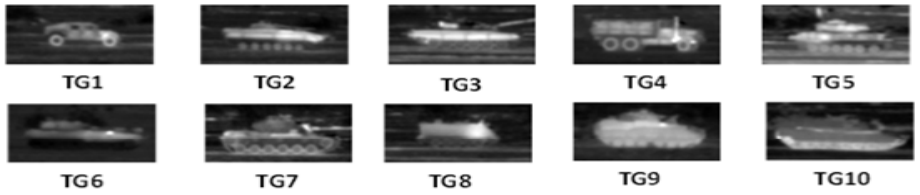

Figure 5: Ten different target vehicles in Comanche IR dataset.

Table 2: Comparison of the proposed method with the recent approaches in terms of recognition rate(%).

	Proposed RGV		SRC methods [14]		
	C.S.	I.G.	Lasso	BPDN	BP
TRAIN-SIG	99.92	99.98	-	-	-
TEST-SIG	99.10	99.10	97.96	97.42	97.69

* RGV is for *K=7, L=7* from Table 1.

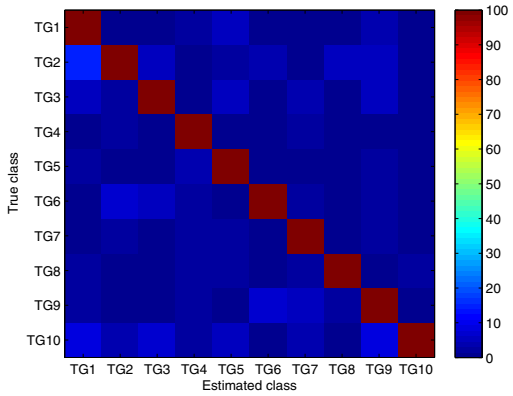

Figure 6: Confusion matrix corresponding to RGV based classification (with vocabulary tree parameter $K = 6, L = 6$ and grouping with C.S.).

We use different vocabulary sizes under different branch factors (K) and tree levels (L). Table 1 shows that the proposed (RGV) outperforms the original BoW model. To form RGV groups, we chose ϵ_t as $\sim 2 \times 10^{-3}$ and $\sim 9 \times 10^{-5}$ for C.S. and I.G. based grouping respectively and ϵ_θ as 0.8. As seen from the table, $\|V\|$ (i.e., the number of visual words) and $\|\Pi\|$ (i.e., the number of RGV groups) are inversely related. When $\|V\|$ is small, then each visual word tends to be a collection of large number of features, hence there will be very little similarity among groups resulting in large $\|\Pi\|$. As $\|V\|$ increases, $\|\Pi\|$ becomes smaller. Therefore, the average number of visual words per group increases from 1.52 to 2099.51 for RGV with I.G. and from 1.44 to 849.35 for RGV with C.S.

Table 2 compares the proposed method with the state-of-the-art SRC based approaches [14]. RGV is able to achieve near perfect results for both training and testing chips which outperforms SRC methods. It is believed that dense HOG features can better encode the thermal distribution of a target signature. A confusion matrix can be useful to assess the efficiency of our proposed method. It represents the number of correct and incorrect predictions (estimated target class) made by the classification model compared to the actual target class (See Figure 6).

3.3. Effect of ϵ_t and ϵ_θ

Forming relevant groups is largely dependent on the choice of ϵ_t and ϵ_θ. Since ϵ_t is found to be more important to create non-overlapping groups, based on our experiments, we only varied the value of ϵ_t and kept ϵ_θ fixed. ϵ_t can be thought as a radius of a circular neighborhood around a visual word v in the feature space and words embraced by this region will form one group. For the sake of simplicity and avoiding the formation of overlapped groups in RGV, words that are once assigned to a group will not

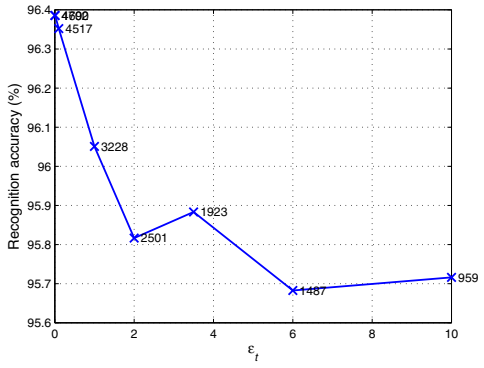

Figure 7: Effect of ϵ_t on recognition performance using *K=7, L=6* and RGV with *C.S.*. Numbers adjacent to points denote $\|\Pi\|$ (number of groups computed at corresponding ϵ_t). $\|\Pi\|$ decreases from 4702 to 959.

be considered for further grouping. A small ϵ_t will produce a large number of groups (large $\|\Pi\|$) with a small average group size (i.e., the number of words per group). On the other hand, a larger ϵ_t can lead to a larger group size with the possibility of grouping non-relevant words together. Eventually, this may introduce ambiguity in Phase-II voting. Thus, when ϵ_t is large enough to contain all words of same dominant class in a single group (*i.e.* 10 groups for 10 target classes), the RGV approach converges to the baseline BoW. As shown in Figure 7, where C.S.-based RGV is implemented with various ϵ_t values for a specific vocabulary (K=7, L=6), ATR recognition accuracy starts from 96.39% with the smallest ϵ_t ($\|\Pi\| = 4702$) and decreases as ϵ_t increases (with smaller and smaller $\|\Pi\|$).

4. Conclusion

In this work, we study the long-standing IR ATR problem and our research is deeply motivated by the recent advancements in the fields of object detection, image classification and retrieval. Based on the idea of BoW, the higher level clustering of similar visual words, constructing visual phrases or semantic categorization are now becoming areas of interest among computer vision researchers. First, it is shown that the dense HOG feature is an effective approach to the ATR task where IR chips usually have limited spatial resolution. Second, the vocabulary tree is useful to reduce the redundancy of dense HOG features by creating a compact set of visual words for scalable object recognition. Third, BoW-based recognition is efficient and robust in the context of IR ATR, although the spatial information of local features are totally ignored. Nevertheless, the introduction of RGV further improves the ATR performance by boosting up the contribution of coexistent visual words belonging to the same relevance group. The reported ATR results on the Comanche IR dataset are among the state-of-the-art in the literature.

Acknowledgment

Part of the computing for this project was performed at the OSU High Performance Computing Center at Oklahoma State University (OSU). The Comanche IR dataset was provided by the Army Research Office (ARO).

References

[1] Spgl1. *http://www.cs.ubc.ca/m̃pf/spgl1*.

[2] *Discovery of collocation patterns: from visual words to visual phrases*, Proc. of CVPR 2007.

[3] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern Information Retrieval, 2nd edition Pages 323-325*. 1999.

[4] B. Bhanu. Automatic target recognition: state of the art survey. *IEEE Transactions on Aerospace and Electronic Systems*, (4):364–379, 1986.

[5] B. Bhanu and T. L. Jones. Image understanding research for automatic target recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 8(10):15–23, 1993.

[6] P. Bharadwaj and L. Carin. Infrared-image classification using hidden markov trees. *IEEE Transactions on PAMI*, 24(10):1394–1398, 2002.

[7] L. A. Chan and N. M. Nasrabadi. An application of wavelet-based vector quantization in target recognition. *International Journal on Artificial Intelligence Tools*, 6(02):165–178, 1997.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of CVPR*, 2005.

[9] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951.

[10] Y. Lamdan and H. J. Wolfson. Geometric hashing: A general and efficient model-based recognition scheme. In *Proc. of ICCV*, 1988.

[11] B. Li, R. Chellappa, Q. Zheng, S. Der, N. Nasrabadi, L. Chan, and L. Wang. Experimental evaluation of FLIR ATR approaches: a comparative study. *Computer Vision and Image Understanding*, 84(1):5–24, 2001.

[12] D. Lin. Automatic retrieval and clustering of similar words. In *In Proc. of the 17th international conference on Computational linguistics 1998*.

[13] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. of CVPR*, 2006.

[14] V. M. Patel, N. M. Nasrabadi, and R. Chellappa. Sparsity-motivated automatic target recognition. *Applied optics*, 50(10):1425–1433, 2011.

[15] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *Proc. of ACM MM*, 2009.

[16] Q. Zhao and J. C. Principe. Support vector machines for sar automatic target recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 2001.