Automatic Classification for Ground Targets under Complex Background Based on Bag of Words Model

Wei Jin\textsuperscript{1,2,a}, Yunsong Feng\textsuperscript{1,2,b}

\textsuperscript{1}State Key Laboratory of Pulsed Power Laser Technology, College of Electronic Countermeasures, National University of Defense Technology, Hefei 230037, China
\textsuperscript{2}Key Laboratory of Infrared and Low Temperature Plasma of Anhui Province, College of Electronic Countermeasures, National University of Defense Technology, Hefei 230037, China

\textsuperscript{a}kingvee\textregistered@163.com, \textsuperscript{b}fyseei\textregistered@163.com

\textbf{Abstract.} In order to improve the performance of the automatic classifier for ground targets under complex background, a new classifier which was based on bag of words model and support vector machine (SVM) was build. Firstly, the keypoints of a target image were extracted into vectors by using Scale-Invariant Feature Transform algorithm, and vectors which were extracted from all trained images were merged into large vectors. Secondly, the large vectors were clustered into small vectors which included $K$ vectors by using $K$-means clustering algorithm. To analyze how different values of $K$ affected the classifying results, $K$ was set to several different numbers. Thirdly, vectors which were extracted from all trained images were classified, and $K$-dimensional numerical vectors were obtained. Finally, the classifier was build with SVM algorithm, and the $K$-dimensional vectors were as input features of the classifier. The training images were randomly chosen from 90% of images dataset which included missile launch vehicles and tanks, and the rest 10% were as the testing images. The trained classifier is tested, and the f1-score value is 0.79. The results show that the new classifier can perform well.

\section{1. Introduction}

The complex background does not have a standard definition, mainly based on a single, simple background. The complex background can generally be understood as: firstly, there are some interference including changeable background, ground objects blocking and shadow; secondly, the quality of images are affected by these factors including changeable sun light, cloudy day, fog and haze; thirdly, the difficulty of detecting will be increased, because the shape and scale of the targets image are changed with the many types of targets, different shooting angles and positions. Automatic detection and classification for ground targets are important for military field. It can be played a key role in knowing dynamic deployment, identification and tracking, precisely striking for vehicle targets including missile launch vehicles, tank and infantry fighting vehicles, etc\cite{1}.

Generally, there are two ways in classifying targets images. Firstly, targets images are matched to the categories templates which can identify the category of target. When the targets are complex and the amount of templates is large, it will cost very large amount of time in calculating\cite{2,3}. Secondly, the images are preprocessed to separating and choosing targets boxes, and extracting target features; the classifier is built by using the machine learning or deep learning methods\cite{4-8}. The second way is
adopted in this paper. The bag of word model is used to automatically classify missile launch vehicle and anti-aircraft gun under complex background.

2. Bag of the word model
The bag of word model was first applied to automatic document classification\(^9\), and then was used to image classification\(^{10-12}\). Similar to document classification, large amount of local invariant features are clustered into small amount of visual words. A image is expressed as a histogram of these visual words. There are four steps to implement the bag of word model: extracting images features, generating a visual dictionary, building a histogram, and creating a classifier.

2.1 Extracting images features
The image features include visual features, global invariant features and local invariant features. Using local invariant features to identify and classify targets in a complex environment can better solve background noise and local occlusion\(^{13}\). Local invariant features include corner points, SIFT (scale-invariant feature transform) features, and contour descriptions.

SIFT which first proposed by Lowe.D in 2004 is the most commonly used image local feature descriptor\(^{14}\). SIFT has rotational invariance, scale invariance and illumination invariance. Therefore, it may eliminate the effects of noise, affine transformation, viewing angle changes. SIFT is adopted to extract images features in this paper.

The SIFT feature extraction algorithm is mainly divided into four steps: scale space extreme value detection, key point positioning, specifying direction parameters and key point descriptors. Through the above four steps, an image is converted into a numerical matrix

\[
m_i = \begin{pmatrix}
I_{1,1} & I_{1,2} & \cdots & I_{1,127} & I_{1,128} \\
I_{2,1} & I_{2,2} & \cdots & I_{2,127} & I_{2,128} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
I_{n-1,1} & I_{n-1,2} & \cdots & I_{n-1,127} & I_{n-1,128} \\
I_{n,1} & I_{n,2} & \cdots & I_{n,127} & I_{n,128} \\
\end{pmatrix}
\]

\hspace{1cm} (1)

\(m_i\) donates the \(i\)th image, and \(n\) donates the amount of key point descriptors. A key point descriptor is described by a 128-dimensional numerical vector.

SIFT features of a missile launch vehicle and an anti-aircraft gun image are extracted and shown in Fig. 1. One circle donates a SIFT feature.

![Fig. 1 SIFT features of a missile launch vehicle and an anti-aircraft gun image](image)

When the SIFT features of all images have been extracted, all features numerical vectors are concatenated into a bigger matrix
2.2 Generating a visual dictionary

The process of generating a visual dictionary is the process of clustering feature vectors. The center of each cluster is regarded as a visual word, and all visual words constitute a visual dictionary. K-means clustering algorithm is highly efficient and is applied frequently for clustering large-scale data.

K-means clustering uses K as a parameter, and divides N samples into K clusters. It makes samples in the same cluster have higher similarity, and samples in the different clusters have lower similarity. The process is in the following.

(1) Setting the number of clusters K. K is a natural number. To compare the effects of different K values on results of classification, K is taken a value of 5 every interval from 10 to 50, and K = (10, 15, 20, ...., 45, 50).

(2) Calculating the center of each cluster. The center of a cluster is updated by calculating the means of samples in the cluster:

\[ \mu_i = \frac{1}{N_i} \sum_{p_i \in C_i} p_i \]

(3) Calculating clustering error. The Euclidean distances between samples of each cluster and the corresponding cluster center are all added, which is used as an error criterion function:

\[ J = \sum_{i=1}^{K} \sum_{p_j \in C_i} |p_j - \mu_i|^2 \]

(4) Determine converge or not. An expected error \( \delta \) is set firstly. If \( J < \delta \), the clustering is over; else repeating steps (1) ~ (4) until \( J < \delta \).

K 128-dimensional numerical vectors which represents the center of each cluster constitute a visual dictionary, such as:

\[
D = \begin{pmatrix}
\mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,127} & \mu_{1,128} \\
\mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,127} & \mu_{2,128} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\mu_{K-1,1} & \mu_{K-1,2} & \cdots & \mu_{K-1,127} & \mu_{K-1,128} \\
\mu_{K,1} & \mu_{K,2} & \cdots & \mu_{K,127} & \mu_{K,128}
\end{pmatrix}
\]

2.3 Building a histogram

All SIFT features of an image are mapped to the closest visual words, and how many times all features of an image appear on each visual word is counted. k-Nearest Neighbor (kNN) classification algorithm is used to solve the problem.
The words of the visual dictionary \( D \) are regarded as samples of which categories are 1 to \( K \) respectively. Suppose an image has \( n \) SIFT features. If a feature belongs to a category by using \( k \)NN the times corresponding to the category will be increased 1. It can be expressed as a vector:

\[
h = (h_1, h_2, \cdots, h_{K-1}, h_K)
\]

The sum of all elements inside the vector \( h \) equals to \( n \). When all SIFT features of \( N \) images are classified by \( k \)NN, it can be expressed:

\[
H = \begin{pmatrix}
h_{1,1} & h_{1,2} & \cdots & h_{1,K-1} & h_{1,K} \\
h_{2,1} & h_{2,2} & \cdots & h_{2,K-1} & h_{2,K} \\
\vdots & \vdots & & \vdots & \vdots \\
h_{N-1,1} & h_{N-1,2} & \cdots & h_{N-1,K-1} & h_{N-1,K} \\
h_{N,1} & h_{N,2} & \cdots & h_{N,K-1} & h_{N,K}
\end{pmatrix}
\]

2.4 Creating a classifier

The support vector machine (SVM) classification algorithm is the most commonly used classification algorithm in supervised learning. Its advantage is that it can obtain better classification results without many samples, and can deal with the problem that the samples are linearly inseparable. The basic idea of the algorithm is to establish an optimal decision hyperplane, so that the distance between the two types of samples on the two sides of the plane from the plane is maximized, thus providing a good generalization ability for the classification problem.

For the two-category problem, after some preliminary derivation, the key to the support vector machine is to solve a constrained optimization problem, which is expressed as follows:

\[
\begin{aligned}
\min_{\omega,b,\xi} & \quad \frac{1}{2} \omega^T \omega + C \sum_{i=1}^{l} \xi_i \\
\text{s.t.} & \quad y_i (\omega^T \phi(x_i) + b) \geq 1 - \xi_i \\
& \quad \xi_i \geq 0, i = 1, \ldots, l
\end{aligned}
\]

The meaning of each variable in the formula can be referred to [15].

3. Experiments

3.1 Establishing targets images database

Two ground targets, missile launchers and anti-aircraft guns, were collected from the open Internet, numbering 447 and 501 respectively, and some images are shown in Fig.2. It is easy to see from the images that the background environment of these targets are more complicated, including: sky, buildings, trees and other military targets. In addition, the targets in the images are also different, including: inconsistent models, different shooting angles, and different lighting. These factors will affect the results of the target classification.

(a) Missile launchers
3.2 Experimental conditions

The program is written in Python. K-means clustering, k-Nearest Neighbor and SVM are implemented by calling the \textit{sklearn} package respectively, and SIFT features are extracted by using \textit{OpenCv} package. All experiments were conducted in a Windows7 PC with an Intel Cerelon G1630 2.80 GHz processor and 4.0G RAM.

3.3 Results

When the experiments begin, the validation dataset is randomly chosen from 10% of the entire dataset. When \(K\) equals to 20, the results are shown in Table 1. the average value of precision, recall and \(f_1\)-score equal to 0.79, 0.78 and 0.79 respectively.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
\textbf{K} & \textbf{precision} & \textbf{recall} & \textbf{f1-score} & \textbf{support} \\
\hline
0 & 0.68 & 0.72 & 0.70 & 35 \\
1 & 0.84 & 0.82 & 0.83 & 66 \\
\hline
avg / total & 0.79 & 0.78 & 0.79 & 101 \\
\hline
\end{tabular}
\caption{Results when \(K=20\)}
\end{table}

To compare the effects of different \(K\) values on results of classification, when \(K\) get value from (10,15,20,……,45,50) in turn, the change of \(f_1\)-score value is shown in Fig.3. The maximum value of \(f_1\)-score is 0.79 when \(K=20\), the minimum value is 0.66 when \(K=35\).

4. Conclusion

In this paper, the bag of word model was used to automatically classify missile launch vehicle and anti-aircraft gun under complex background. Firstly, SIFT features of target images were extracted. Secondly, SIFT features were clustered supervised with K-means clustering algorithm. Thirdly, a histogram was obtained with k-Nearest Neighbor. Finally, histograms of all targets images were classified with SVM. When \(K=20\), the classification results were the best, and \(f_1\)-score = 0.79. The results show that the bag of word model can perform well when it is used to automatically classify ground targets under complex background.
References

[1] XIE Xiaozhu, HE Cheng. Review of vehicle recognition in complex environment background [J]. Journal of Ordnance Equipment Engineering, 2017, 38(6): 90-94.

[2] Zhu Xufeng, Ma Caiwen, Liu Bo. Aerial target automatic classification based on improving bag of words model [J]. Journal of Infrared and Laser Engineering, 2012, 41(5): 1384-1388.

[3] LIN Yuchi, CUI Yanping, HUANG Yinguo. Study on edge detection and target recognition in complex background [J]. Journal of Optics and Precision Engineering, 2006, 14(3): 509-514.

[4] LIU Tianyu, FENG Quan. Detecting grape leaves based on convolutional neural network [J]. Journal of Northwest University (Natural Science Edition), 2017, 41(4): 505-512.

[5] WANG Hong, MO Yumeng, YIN Lijun, et al. An image recognition algorithm on small arms shooting target in complex field background [J]. Journal of Fire Control& Command Control, 2009, 34(7): 60-62.

[6] DENG Liu, WANG Zijie. Deep convolution neural networks for vehicle classification [J]. Application Research of Computers, 2016, 33(3):930-932.

[7] ZHANG Feiyun. Car detection and vehicle type classification based on deep learning [D]. Bengjiang: Jiangsu University, 2016.

[8] CAI Yingfeng, WANG Hai, CHEN Long, et al. Robust vehicle recognition algorithm using visual saliency and deep convolutional neural networks [J]. Journal of Jiangsu University (Natural Science Edition), 2015, 36(5): 331-336.

[9] YU Minjie. Bayesian network classifier and applications [D]. Kunming: Yunnan University of Finance and Economics, 2012.

[10] Fei-Fei L, Perona P. A Bayesian hierarchical model for learning natural scene categories [C] // Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005: 524 – 531.

[11] Roberto Toldo, Umberto Castellani, Andrea Fusiello. The bag of words approach for retrieval and categorization of 3D objects [J]. The Visual Computer: International Journal of Computer Graphics-Special Issue on 3D Object Retrieval, 2010, 26(10): 1257-1268.

[12] XIAO Zhe, QIN Zhiguang, DING Yi, et al. Efficient method for image classification based on low-scale bag of word model [J]. 2016, 45(6): 997-1001.

[13] CAO Jian, CHEN Hongqian, MAO Dianhui, et al. Survey of image object recognition based on local features [J]. Journal of Central South University (Science and Technology), 2013, 44(2): 258-262. (in Chinese)

[14] LOWE D. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[15] C.-C. Chang, C.-J. Lin. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27:1-27:27.