www.s2is.org
International Journal on Smart Sensing and Intelligent Systems
ISSN 1178-5608

A STUDY OF X-RAY MACHINE IMAGE LOCAL SEMANTIC FEATURES EXTRACTION MODEL BASED ON BAG-OF-WORDS FOR AIRPORT SECURITY

Ning Zhang [1] and Jinfu Zhu [*2]
1. College of Civil Aviation of Nanjing University of Aeronautics & Astronautics, Nanjing, China
College of Civil Aviation of Guangzhou,China
2. College of Civil Aviation of Nanjing University of Aeronautics & Astronautics, Nanjing, China

Submitted: Aug. 25, 2014 ***Accepted: Jan. 7, 2015*** ***Published: Mar. 1, 2015***

Abstract- The aviation security at the airport has been faced with increasingly severe situations since the 9-11 event. It's of utmost importance to train airport X-ray machine screener's image recognition competency. So they can prevent terrorists from bringing dangerous articles in their carry-on or checked bags. However, usually the luggages are placed in different positions and the density & volume of articles differ greatly. As a result, dangerous articles show a variety of X-ray image features. It's easy for the confused screeners to miss or incorrectly detect dangerous articles. This has been a hidden danger for civil aviation safety. For image recognition improvement, the researcher analyzed the visual semantics of dangerous goods images and applied a local semantic features extraction method. After classification and summarization, the method was used to train the screeners for particular image recognition. The comparison showed the improved accuracy and efficiency of image recognition for the screeners and demonstrated a satisfactory effect.

Index terms*:* airport security, bag-of-words, semantic analysis, image features

I. INTRODUCTION

The daily job of airport X-ray machine screeners is very busy and tough. ICAO (International Civil Aviation Organization) generally requires that screeners should recognize and analyze each X-ray machine image within 8 seconds. If any dangerous goods images are detected, the luggage should be opened for further check. But usually the luggage are placed at different angles and locations on the conveyer belt. The articles have different density and volume. Consequently the two-dimensional X-ray images shown on the computer screen vary from each other. Then it's more difficult for screeners to accurately and effectively recognize the dangerous article images. Figure 1 shows the two-dimensional images of the same pistol placed at three different positions in the luggage. The image B is easy to be recognized as a pistol by the screeners. But the image A and C, produced when the pistol is revolved horizontally and vertically, cannot easily show the shape of the pistol. Therefore screeners easily make wrong judgments. (As shown in Figure 1)

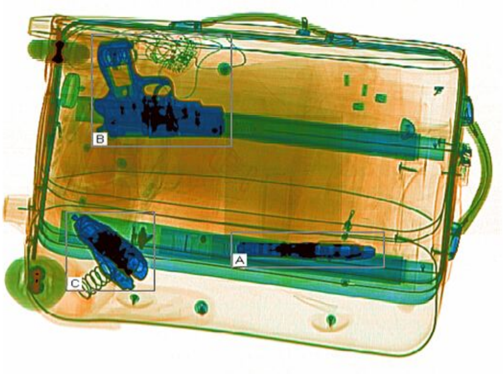


Figure 1. Different images on the X-ray screen for a pistol placed at different angles in the luggage.

In 2004, Maneesha Singh put forward a knowledge-based framework for X-ray image enhancement in his research, but he didn't generalize or classify the image features [1]. In 2007, Stefan Michel carried out the four-month training and research for screeners by using the computer-based training (CBT) system. He proposed to increase image recognition competency with the help of the computer system [2]. In 2009, Saskia M. Steiner, in his paper entitled "Assessment of X-Ray Image Interpretation Competency of Aviation Security Screeners", presented the concept of X-ray screening classification. He only simply classified the goods and then recognized the images. He didn't analyze the image features. However, he paid attention to the cultivation of X-ray image interpretation competency for screeners [3]. In his research of

2010, Claudia C. von Bastian focused on the importance of color recognition for X-ray screening [4]. In 2010, Giby Raphael suggested that the interactive neuro-educational technologies (I-NET) were helpful to the image recognition training. It was a method to explore the image recognition training from the psychological perspective [5]. In 2011, Marcia Mendes published a paper about the influence of virtually-merged images on the effectiveness of computer-based training in X-ray screening. And that was the beginning of simple image classification researches [6]. In 2009, Stefan Michel and Adrian Schwaninger studied the image recognition improvement from the perspective of human-machine interaction on the basis of computer-based training (CBT) [7]. In 2009, Anton Bolfin and Adrian Schwaninger carried out a research on the selection and pre-employment assessment in aviation security X-ray screening. They studied the factors influencing image recognition efficiency, such as color, angle and density [8]. In 2013, Stefan Michel and other scholars put forward the SURF(Speed Up Robust Features)-based bag-of-word(BoW) image recognition method and the concept of bag-of-visual-words (BoVW) [9].

So far, most of the researches on airport X-ray machine images have focused on the image sharpness, color and gray level. They do not systematically classify and generalize the image features of dangerous articles. The above-mentioned researches are contributive to some extent in the aspect of airport X-ray image recognition, but they haven’t classified the images in detail or formed a systematic bag of visual words. In this paper, the X-ray dangerous article images are classified and sorted out. The visual words semantics are analyzed. A bag of visual words for X-ray images and a Bag-of-Words model are established for image feature extraction and systematic classification. The research will provide scientific instructions for the regular X-ray machine training. Consequently the image recognition training will be specific to improve the check efficiency and accuracy. The probability of dangerous articles being missed or judged incorrectly will decrease.

II. METHODS

2.1 Image Semantics

Image semantics[10], namely the interpretation of the image contents, is a concept referring to the process where the information is extracted from the essential attributes of the images and the low-level information is transmitted, mapped and integrated into the high-level semantics. The image semantics extraction is generally actualized through the mapping between the low-level features and the high-level semantics. The existing mainstream image semantics researches construct the semantic model based on systematic learning. The model obtains the low-level visual features of images and establishes a one-way mapping relationship between visual features and semantics through pattern recognition and machine learning. This paper, on the basis of the Bag-of-Visual-Words model and without supervision, establishes the semantic model from low-level visual features to high-level semantic features. It's established through the local semantic features of the visual words of dangerous article images.

2.2 Visual Image Features

2.2.1 Definition of Visual Image Features

Visual image features[11], as the attribute data which can represent the semantics of the images, represent the images themselves and will be used in the subsequent analysis and process. Since the image contents are described through their features and influence the subsequent comparison of similarity, the selection of visual features is particularly important. The researcher summarized the visual images of the same pistol placed at different angles in the luggage. (As shown in Figure 2).

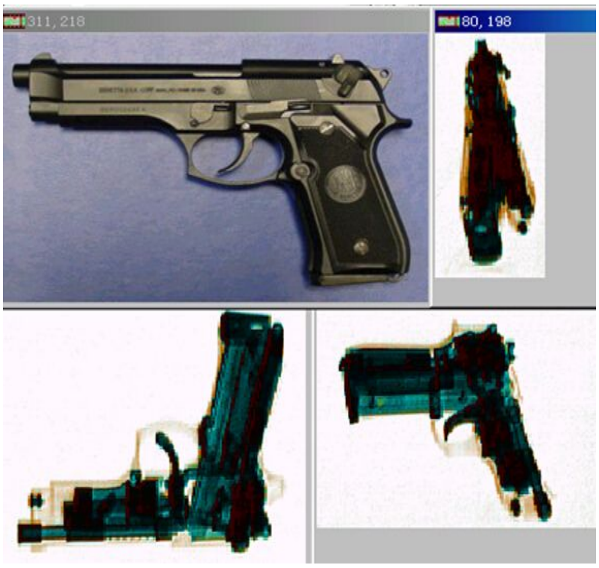


Figure 2. The visual words features of the pistol on the X-ray machine screen

22.2. Levels of Visual Image Features

The semantics of image scenes are characterized by levels. According to the different levels of image understanding, the images can be roughly categorized into three levels, the low level, middle level and high level respectively. The low level, as the feature level, is mainly characterized by the low-level visual features, such as color, texture, shape and edge features. The middle level refers to the intermediate semantic features obtained through modeling and deduction on the basis of the low-level feature extraction. The middle-level semantic modeling methods include semantic object method, semantic attribute method and local semantic concept representation method. The high level is the scene semantic concept obtained through the higher level abstraction of the images. Among the three levels of scene semantics, the upper level is more abstract and advanced than the lower one; the high-level semantic concept can be inferred by the low-level features.

1) Low-Level Visual Image Features

The low level shows the image scenes with the low-level features, which is the early common method for image scene classification and retrieval technology. The low-level features can be acquired directly from the image data and can objectively reflect the image contents. Consequently the low-level feature extraction is the basis for image scene classification. The low-level features of the images are consisted of the global and local features. The commonly used global features include color, texture, edge and shape, etc.

Color is a kind of important visual information. As the basic feature of images, color is also the focus of attention among the traditional image classification and retrieval technology. The color feature has the least dependency on translation, rotation, scale change and even various transformations of images and therefore demonstrates the extraordinarily strong robustness. The color feature is less dependent on the size, direction and perspective of the image and therefore it’s more robust.

Texture is a visual feature reflecting the homogeneity phenomenon of images. It does not depend on color or brightness. It is the intrinsic characteristic of all objects surfaces. Essentially the texture feature is the law to depict the pixel gray level distribution in the neighborhood.

Shape is another important low-level visual feature of images. It provides a means to describe the high-level visual features and plays an important role in further acquiring image semantics. The shape description methods are classified into two: one is based on the boundary and the other is based on the region. The former only utilizes the outer boundary of the shape, such as the Fourier descriptor and skeleton description; while the latter uses the whole shape region, such as invariant moments and the region area. The global feature extraction process is simple and easy for calculation, but disadvantageously they can only reflect the global statistical information of images and neglect the local detailed information. As researches have shown, the local detailed information of images tends to embody the nature of images.

2) Middle-Level Visual Image Features

The middle-level semantic modeling is put forward to reduce the “semantic gap[12]” between the low level and high level. It divides the bigger “semantic gap” between the low-level feature layer and the high-level feature layer into two smaller “semantic gaps”, that is, the gap between the low-level feature and the middle-level semantic concept and the gap between the middle-level semantic concept and the high-level scene semantics.

The middle-level semantic modeling methods include semantic object method, semantic attribute method and local semantic concept representation method. The semantic object method describes the whole image through detecting or recognizing the target object of images; the semantic attribute method describes the image scene according to the predefined semantic attributes and represents the image semantics by use of sensory attribute concepts. The local semantic concept representation method constructs the mapping from local features to local semantic concepts and then represents the scene semantics in accordance with the local semantic concept distribution among images.

The local semantic concept representation method is the currently most widely used one of middle-level semantic modeling. In light of different mapping ways, the local semantic concept representation methods can be classified into three kinds: the first based on the Bag-of-Visual-Words model, the second based on the sparse coding model and the last based on the semantic topic model.

a. By referring to the idea of the Bag-of-Words model in the text processing field, the visual words model regards the image as a “document” consisted of visual words. It then computes the frequency of occurrence of the visual words to construct a frequency distribution histogram representation of the image.

b. The sparse coding model generates a visual vocabulary and a sparse coefficient matrix for the image by using the sparse coding theory to construct a sparse vector representation of the image.

c. The semantic topic model learns the semantic topic in the image scene and constructs a semantic topic representation of the image on the basis of the visual words distribution and by using the statistical model or probabilistic generative model.

3) High-Level Visual Image Features

The high level refers to the high-level semantic information of the image comprehended in the human cognitive style. It includes scene semantics, behavioral semantics and emotional semantics, etc. The scene semantics, such as the beach, city and forest, is the representation of image contents and can be used in scene classification and image retrieval. The behavioral semantics, such as the tug of war or football match, represents the more abstract event information in the image. The emotional semantics, such as anger, excitement and fear, represents the emotional information embodied in the image

2.3Summarization of the Image Scene Hierarchical Model

The researcher summarized the image scene hierarchical model as the following: (illustrated by Figure 3)

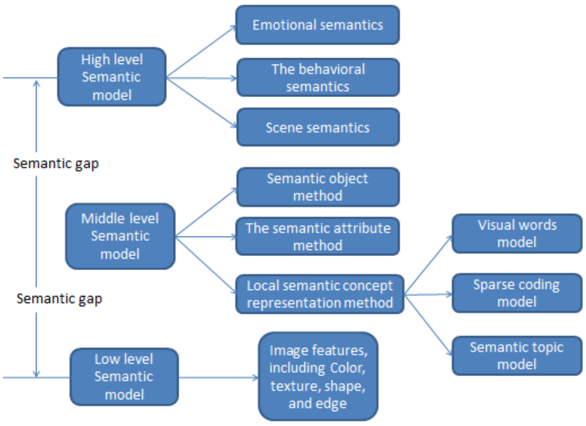


Figure 3. Flow chart of the hierarchical model of the image scene

2.4Image Scene Classification Methods Based on Low-Level Image Features

According to the different levels of describing image scene contents, there are two types of image scene classification methods: one based on low level features and the other based on middle level semantic modeling.

The early scene classification researches are based on low level image features. They first use low level features, such as color, texture and shape, to establish the image scene description and then deduces the high level semantic information of the image scene by adopting classifiers. The principle of low-level-features-based image scene classification method originates from the semantic categories of image scenes. They can be discriminated directly through low level visual features without any need to identify the specific objects in the image. As for this sort of scenes, the gap between low level features and high level semantics can be bridged on the basis of low-level image feature extraction combined with the machine learning method. Nevertheless, this method is only applicable to the simple image scene classification and not suitable for complex image scene classification tasks.

2.5 Image Scene Classification Method Based on Local Semantics

High level semantics acquisition is of utmost importance to successful image analysis and understanding. The construction of the probability relation between low level visual features and image semantics reflects the conversion between data and concepts. The researches based on image semantics focus on analyzing, extracting and describing the middle and low level visual data features and modeling the mapping relationship between them and local feature semantics.

The middle level semantic representation is actually to construct the mapping between low level image features and middle level semantic concepts to cross the semantic gap between the low level and the middle level; concurrently the image scenes of the same category usually show similar semantic concept distribution. So the image scenes can be classified, through the middle level semantic representation and in accordance with the semantic concept distribution, to bridge the semantic gap between the middle level and high level.

This paper studied the scene classification method based on local semantic concept representation. The basic philosophy of the semantic topic modeling of image scenes is: the image scene is first considered to be consisted of multiple semantic topics; then the semantic topic distribution of the image is learned through the probabilistic generative model and finally the image scene categories are determined in light of the semantic topic distribution. The image scene classification method based on local semantic concept representation constructs the middle level semantic representation, achieves the mapping between low level visual features and middle level semantic concepts and ultimately completes the image scene classification on the basis of semantics.

In reality, the X-ray images of dangerous articles are the abstract representations of their shapes, pictographic but always incomplete. They can be regarded as the visual representation of the local semantics for dangerous goods images. The semantic feature extraction method used in this paper is fully applicable to the abstract image interpretation of dangerous goods in baggage security checkpoints.

2.6 Research on Image Feature Extraction Classification Algorithm Based on Local Semantics

In different visual tasks, the image can be described in different feature attributes (color, brightness, outline and shape), or multiple feature image representation can be generated via feature fusion technology. *P* indicates the feature attribute set: $P=\{p_1,p_2,\dots,p_m\}$, $p_i(i=1,2,\dots,m)$ refers to the attribute *i* in the attribute set. If each attribute is described with the *n*-dimensional feature vector, a low-level *m*×*n*-dimensional feature space *X* will be generated. In the feature space, the extracted local image feature *x* can be described with the feature vector as:

$$x=\left(x_1^1,x_2^1,\cdots,x_n^1;x_1^2,\cdots,x_n^2;\cdots;x_1^m,x_2^m,\cdots,x_n^m\right) \tag{1}$$

The feature vector is a digital description of images. Its set is used to replace the original image, namely the feature set of the original image. Then the image I can be consisted of the point sets in the feature space in accordance with the particular distribution: $I\approx X=\{x_i\}$. x_i represents the feature *i* obtained from the formula (1). Generally in practice, only one feature attribute is selected for description, that is, *m*=1, then the feature vector can be simplified as $x = (x_1^1,\ x_2^1,\dots,\ x_n)$.

In this paper, the image semantic space is represented by $Z = \{z_1, z_2,\dots,z_k\}$ when the image attribute is extracted to create semantics. Z_i = （$i=1,2,\dots,k$） indicates the semantic concept *i* consisting the *k*-dimensional semantic space. As one of the existing image semantic researches, this research constructs the semantic model based on systematic knowledge. The core of the model is to build a one-way mapping relationship between visual features and semantics. Then the mapping from the low-level *n*-dimensional feature space *x* to the *k*-dimensional semantic space *z* is achieved: $f:X^{n}\rightarrow Z^{k}$

The image semantic space is represented by

$Z=\{z_1,z_2,\dots,z_k\}$,$z_i(i=1,2,\dots,k)$indicates the semantic concept *i* consisting the *k*-dimensional semantic space.

The image set is represented by $D = \{d_1, d_2,..., d_D\}$; the bag of visual words is represented by $W = \{w_1, w_2,..., w_N\}$;

$N = (n_{ij})_{D\times N}$ refers to the co-occurrence frequency matrix of the images and words; $n(d_i,w_j)(i=1,2,...,D,j=1,2,...,N)$ indicates the frequency of the word w_j occurs in the document d_i; $Z=\{z_1,z_2,...,z_k\}$ represents the set of latent semantics. Probabilistic Latent Semantic Analysis (PLSA) assumes that images and words are conditionally independent, and the latent semantics is also conditionally independent in the image or word distribution. Based on the above assumption, the formula (2) can be used to represent the joint probability of visual images and visual words:

$$P(d_i,w_j)=P(d_i)P(w_j|d_i)= P(d_i)\sum_{k=1}^{K}P(z_k,w_j|d_i)= P(d_i)\sum_{k=1}^{K}\frac{P(z_k,w_j,d_i)}{P(d_i)}= P(d_i)\sum_{k=1}^{K}\frac{P(z_k)P(w_j,d_i|z_k)}{P(d_i)}= P(d_i)\sum_{k=1}^{K}P(w_i|z_k)P(z_k|d_i) \quad (2)$$

In the formula, $P(w_j|z_k)$ refers to the distribution probability of latent semantics in words and also the semantic probability distribution embodied by the words corresponding to the image local features. $P(z_k|d_i)$ indicates the distribution probability of latent semantics in images. Both are the very key parameters for image semantic extraction and the ultimate goals of the semantic model. The expectation maximization algorithm (Expectation Maximization, EM) can be used to solve the equation.

The maximum likelihood function, formula (3), is constructed on the basis of formula (2). Computing the optimal distributions of parameters $P(w_j\ |z_k)$ and $P(z_j\ |d_i)$ is equivalent to computing the maximum value of likelihood probability $P(d_i,w_j)$:

$$\max l = \lg\prod_{i,j}P\left(d_i,w_j\right)^{n\left(d_i,w_j\right)} \quad (3)$$

The formula (3) is converted into the maximum

value problem of the log-likelihood function and the exponent arithmetic is converted into the multiplication:

$$L = \lg\prod_{i,j}P\left(d_i,w_j\right)^{n\left(d_i,w_j\right)} = \sum_{i,j}\left(d_i,w_j\right)\lg P\left(d_i,w_j\right) \quad (4)$$

The maximized L:

$$\max L = \max\sum_{i=1}^{D}\sum_{i=1}^{N}n\left(d_i,w_j\right)\lg P\left(d_i,w_j\right) = \max\sum_{i=1}^{D}\sum_{i=1}^{N}n\left(d_i,w_j\right)\lg P\left(d_i\right)P\left(w_j|d_i\right) =$$

$$\max\sum_{i=1}^{D}\sum_{i=1}^{N}n\left(d_i,w_j\right)\left[\lg P\left(d_i\right)+\lg P\left(w_j\left|d_i\right.\right)\right]= \max\sum_{i=1}^{D}\sum_{i=1}^{N}n\left(d_i,w_j\right)\lg P\left(d_i\right)+$$
$$\sum_{i=1}^{D}\sum_{i=1}^{N}n\left(d_i,w_j\right)\lg P\left(w_j\left|d_i\right.\right)= \max\sum_{i=1}^{D}\sum_{i=1}^{N}n\left(d_i,w_j\right)\lg P\left(w_j\left|d_i\right.\right) \tag{5}$$

As in $P(w_i|d_i)=\dfrac{P(w_i,z_k|d_i)}{P(z_k|w_j,d_i)}$

$P(w_i,z_k \mid d_i)$is the full parameter, while $P(w_i \mid d_i)$ is the incomplete parameter. The default parameter in

$P(w_i \mid d_i)$is computed by the EM algorithm. The expectation step (M-step) is:

$$\sum_{i=1}^{D}\sum_{j=1}^{N}n\left(d_i,w_j\right)E\left[\lg P\left(w_j\left|d_i\right.\right)\right]_{z_{k|w_j},d_i} = \sum_{i=1}^{D}\sum_{j=1}^{N}n\left(d_i,w_j\right)E\left[\lg P\left(w_i,z_k\left|d_i\right.\right)-\lg P\left(z_k\left|w_j,d_i\right.\right)\right]_{z_{k|w_j},d_i} =$$
$$\sum_{i=1}^{D}\sum_{j=1}^{N}n\left(d_i,w_j\right)E\left[\lg P\left(w_j\left|d_i\right.\right)\right]\lg\left(w_i,z_k\left|d_i\right.\right) \tag{6}$$

$$P\left(z_k\left|d_i,w_j\right.\right)=\frac{P\left(w_j\left|z_k\right.\right)P\left(z_k\left|d_i\right.\right)}{\sum_{l=1}^{K}P\left(w_i\left|z_l\right.\right)P\left(z_l\left|d_i\right.\right)} \tag{7}$$

Under the constraint condition:

$$\sum_k P(z_k|d_i)=1$$

$$\sum_j P(w_j|z_k)=1$$

The optimized formula (7), the Lagrange multiplier method, is applied together with formula (6). The maximum parameter estimation equation (M-step) based on EM parameter solution is as follows:

$$P\left(w_i\left|z_k\right.\right)=\frac{\sum_{i=1}^{D}n\left(d_i,w_j\right)P\left(z_k\left|d_i,w_j\right.\right)}{\sum_{m=1}^{N}\sum_{i=1}^{D}n\left(d_i,w_m\right)P\left(z_k\left|d_i,w_m\right.\right)}$$

$$P\left(z_k \middle| d_i\right) = \frac{\sum_{i=1}^{D} n\left(d_i, w_j\right) P\left(z_k \middle| d_i, w_j\right)}{\sum_{m=1}^{K} \sum_{j=1}^{N} n\left(d_i, w_j\right) P\left(z_l \middle| d_i, w_j\right)} \tag{8}$$

In the semantic extraction parameter training stage, the above models are applied to all the image local feature sets generated by training images. *P(w|z)* is finally obtained through the EM algorithm iteration until convergence. *P(w|z)* is actually the latent semantic model of local features. It describes the distribution law of visual words when the latent semantics in the image local features appears. In the inference stage, *P(w|z)* remains unchanged for all local features of the test images. *P(z|d)* of each image is finally obtained through the EM algorithm iteration until convergence. *P*(*z*|*d*) represents the probability distribution of images in the semantics z. Based on this can the image contents be analyzed and understood.

P(*z*|*w*) can be gained from the Bayes formula, and it reflects the distribution of image local features in semantics. Assuming that the number of latent semantics is *K*, a *K*-dimensional feature vector $(P(z_1|w_j),...P(z_k|w_j))$ can be gained for each local feature w_j. If the feature number of an image is *M,* a *M*×*K*-dimensional semantic feature vector $(P(z_1 | w_1),...P(z_k | w_1),..., P(z_1 | w_M)..., P(z_k | w_M))$ will be obtained. After the image local feature latent semantic features are acquired, a SVM classifier model can be constructed to classify the image scenes. The images are optimized according to the new classification methods and a bag of visual words for sharp weapons is produced, as illustrated by Figure 4.

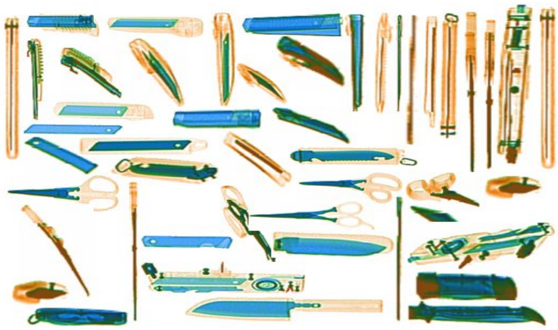


Figure 4. Image combination of various sharp weapons on the X-ray machine screen in visual words

III. EXPERIMENT ANALYSIS

3.1Test Training Images

The image scenes of 15 dangerous articles, including the pistol, hair gel, alcohol, toxic, TNT and explosives, were selected from the X-ray image visual vocabulary. These dangerous goods were categorized. For each category, 100 scene images were chosen. They were placed in different positions and their volume & density varied. So the images were different from the original appearances. Some were quite difficult to be identified. 100 images of them were selected randomly as the training set and the rest were treated as the test set. In the training set, each category of dangerous goods were trained for the semantic models to find the image set of the typical semantic image representation, $P(z|d)$. Then the Support Vector Machine was adopted for classification. The classification results were analyzed by using the confusion matrix.

The confusion matrix[13, 15-16], as a visualization tool, can describe the relationship between the sample data authenticity and the recognition result type. Assuming a classification task consisted of N categories of patterns. The recognition data set D contains T_0 samples; each pattern includes T_1 data ($i=1, 2, ..., N$). C is the classifier adopting some recognition algorithm; cm_{ij} indicates that the i pattern is judged by the classifier C as the j pattern data's percentage of the total samples for the i pattern data. Finally the $N\times N$-dimensional confusion matrix CM (C,D) is acquired.

$$CM(C,D)=\begin{pmatrix} cm_{11} & cm_{22} & \cdots & cm_{1i} & \cdots & cm_{1N} \\ cm_{21} & cm_{22} & \cdots & cm_{2i} & \cdots & cm_{2N} \\ & \cdots & & & \cdots & \\ cm_{i1} & cm_{i2} & \cdots & cm_{il} & \cdots & cm_{iN} \\ & \cdots & & & \cdots & \\ cm_{N1} & cm_{N2} & \cdots & cm_{Nl} & \cdots & cm_{NN} \end{pmatrix} \tag{9}$$

The row subscript of the confusion matrix elements corresponds to the real attribute of the target. The column subscript corresponds to the recognition attribute generated by the classifier. The diagonal element represents the percentage of each pattern correctly recognized by the classifier C. The non-diagonal element indicates the percentage of error judgment.

The correct recognition rate of the classifier for each pattern can be obtained from formula (9):

$$R_i = cm_{il}, i=1,2,\ldots,N \tag{10}$$

Average correct recognition rate:

$$R_A = \sum_{i=1}^{N} (cm_{ij} T_i) / T_0 \tag{11}$$

Incorrect recognition rate for each pattern:

$$W_i = \sum_{j=1, j\neq i}^{N} cm_{ij} = 1 - cm_{ij} = 1 - R_i \qquad (12)$$

Average incorrect recognition rate:

$$W_A = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \left(cm_{ij} T_i \right) / T_0 = 1 - R_A \qquad (13)$$

The confusion matrix of the categorized dangerous goods image scenes is as illustrated by Table I.

Table I. THE CONFUSION MATRIX OF DIFFERENT DANGEROUS GOODS SCENES

Confusion Matrix	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
C1:TNT	84	5	0	7	0	0	0	0	4	0	0	0	0	0	0
C2: Knives	0	80	7	6	0	0	0	0	0	1	2	2	0	0	2
C3: Bullet	0	7	76	8	1	0	0	0	1	2	0	2	1	0	2
C4: Toxic	6	8	4	77	3	4	0	0	0	5	2	0	2	0	2
C5:Gun	5	8	4	0	88	2	0	0	2	1	0	0	1	0	0
C6:Pistol	0	2	0	0	3	82	0	6	7	0	0	0	0	0	0
C7:Alcochol	9	0	0	0	2	0	77	1	0	0	0	0	0	0	0
C8:Scissors	3	0	2	5	0	0	2	78	3	6	0	0	0	0	0
C9:Hairgel	2	0	0	1	0	0	3	0	89	0	0	1	0	0	2
C10:Tear gas	1	0	2	0	0	0	2	6	2	81	0	8	10	6	1
C11:Crossbow	0	1	0	1	0	0	0	0	2	0	86	7	1	7	2
C12:Detonator	0	0	1	0	2	3	2	0	9	0	1	83	5	3	1
C13:Fireworks	0	0	0	1	0	0	0	0	0	0	2	1	75	2	1
C14:Lighter gas	0	1	0	0	2	0	0	0	3	0	6	3	1	78	2
C15:Expolosives	3	5	7	4	0	0	0	0	0	0	7	4	2	1	91

In the confusion matrix, the x-axis and y-axis represent the scene category respectively. The value in row *i* and column *j* indicates the proportion of the *i* image classified as the *j* image. The diagonal element value in the confusion matrix represents the classification accuracy for each type of scenes. The average classification accuracy is

calculated by formula (10):

$R_A = \sum_{i=1}^{N} (cm_{ij} T_i) / T_0 =$ 79.13%. As shown in Figure 5, the left picture is the visualization confusion matrix; the progress bar in the right picture indicates that the represented classification rate increases with the brightness.

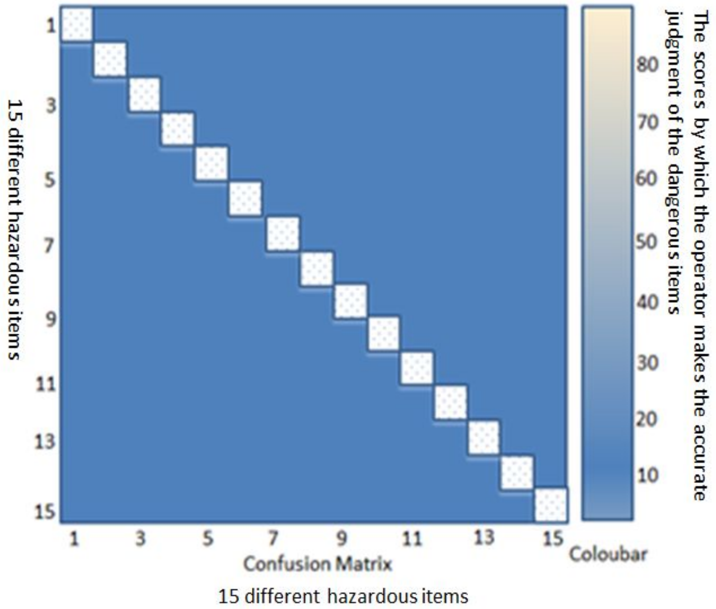


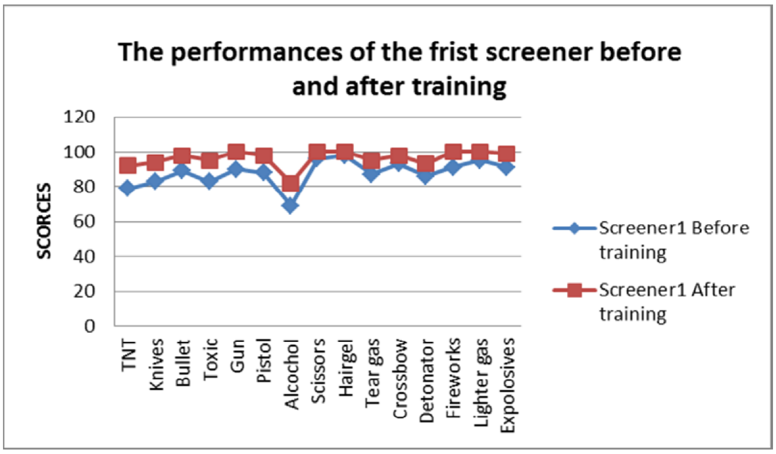
Figure 5. Visualization confusion matrix

Left graph,1 to 5 stands for 15 different hazardous items. Right graph, coloubar, 0 to 100 stands for the percentage by which the operator makes the accurate judgment of the dangerous items.

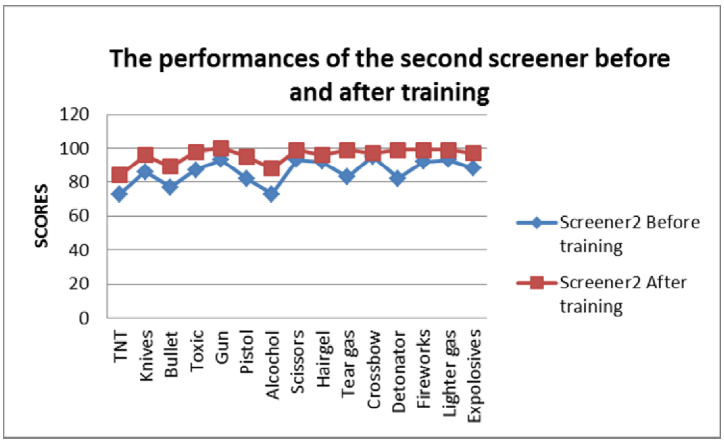
3.2Experiment Training Analysis

The researcher optimized the images according to the new local feature classification and carried out the targeted experiments. Five X-ray machine screeners with the senior security officer qualification from some hub airport in South China were invited. Each of them operated a German Smiths Heiman HI-SCAN 10080 EDS machine. They were trained, according to the given image library after classification, to look for the regular features of images. After one month of training, they were tested to identify these 1500 dangerous goods images within 4 hours. The results were compared with their performances before training. Their competency in recognizing the test images was evaluated. The recognition accuracy for each screener increased at least

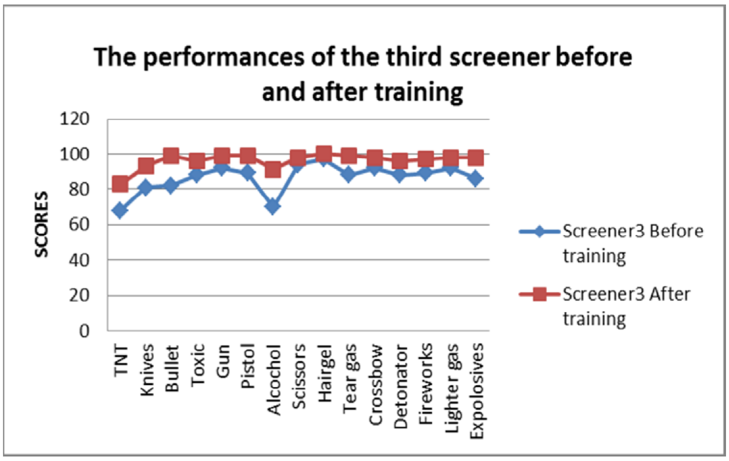
by 9.9%. Some even increased by 12.03%. The average time of check articles decreased by 20.7%. They distinctly improved their ability in identifying TNT[14] explosives. Due to the intensive training in bag of visual words, the 5 screeners gained further understanding of the color, gray level and edge features of the local images of TNT explosives. Their average performance improved by 35.28%. (As shown in Table II, Figure 6and Figure7)



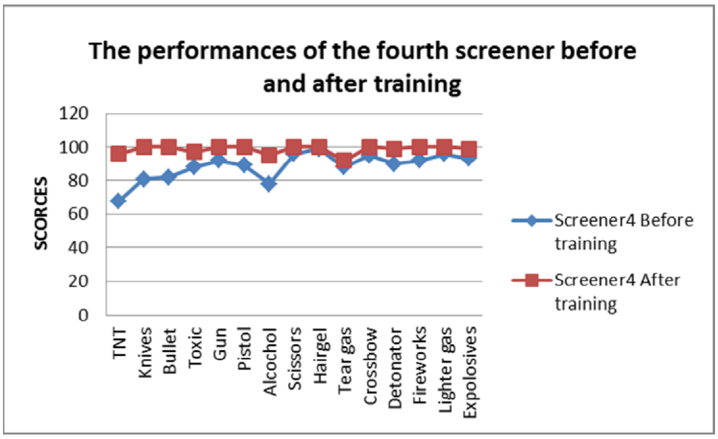
(A) THE PERFORMANCES OF THE FIRST SCREENER BEFORE AND AFTER TRAINING



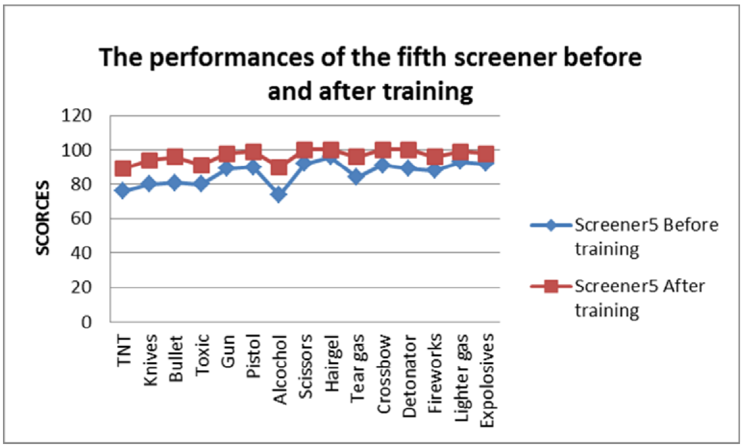
(B) THE PERFORMANCES OF THE SECOND SCREENER BEFORE AND AFTER TRAINING



(c) The performances of the third screener before and after training



(d) The performances of the fourth screener before and after training



(e) The performances of the fifth screener before and after training

Figure 6. The performances of different screeners before and after the image training (a), (b), (c), (d), and (e)

The X axis shows 15 different hazardous items. The Y axis demonstrates the scores that measure the operator’s accurate judgment of hazardous items, and the full score is 100.

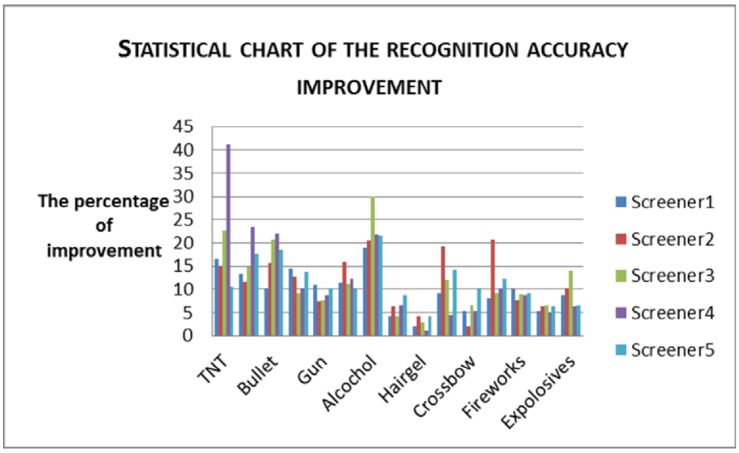


Figure 7. Statistical chart of the recognition accuracy improvement of all screeners before and after training

The X axis shows 15 different hazardous items. The Y axis shows the percentage of improvement between scores operators achieve before and after training.

IV. Conclusion

The traditional semantic image methods of visual scenes require manual marking. The method adopted in this paper explores the local latent semantic features directly from low level image features and without any supervision. It’s an expectation-maximization extraction method for local image semantic features. It achieves the automatic mapping from the image low level to the high-level feature semantics. As a method without supervision, it can be used to compute the probability distribution of local semantics. The image local semantic features can be applied to scene classification to acquire the space distribution features of local semantics. Consequently, the dangerous goods images can be modeled based on scenes and the visual vocabulary for systematic classification can be established. The targeted image recognition training for the screeners demonstrated a good expected effect and proved the applicability and feasibility of this method.

References

[1] Maneesha Singh, Sameer Singh, Derek Partridge. “A Knowledge-Based Framework for Image Enhancement in Aviation Security”, *IEEE Transactions on Sysytems, Man, and Cybernetics—Part B: Cybernetics,* vol. 34, NO. 6, pp. 2354-2365, December 2004.

[2] Stefan Michel, “Computer-Based Training Increase Efficiency in X-ray Image Interpretation by Aviation Security Screeners”, *Security Technology, 2007 41st Annual IEEE International Carnahan Conference on Date of Conference:* 8-11 Oct. 2007, pp. 201 – 206.

[3] Saskia M. Steiner-Koller, Anton Bolfing, Adrian Schwaninger. “Assessment of X-Ray Image Interpretation Competency of Aviation Security Screeners”, *Security Technology, 2009. 43rd Annual 2009 International Carnahan Conference on Date of Conference*: 5-8 Oct. 2009. pp. 20 – 27.

[4] Claudia C. von Bastian, “Colour Impact on Security Screening”, *IEEE A&E Systems Magzine*, pp.1-6, October 2010.

[5] Raphael,G.; Berka,C.; Kintz,N.; Tan,V. “Interactive Neuro-Educational Technologies (I-NET): Enhanced Training of Threat Detection for Airport Luggage Screeners”, *Technologies for Homeland Security (HST), 2010 IEEE International Conference on Date of Conference*:8-10 Nov. 2010. pp:536 – 542.

[6] Marcia Mendes, “Does the Application of Virtually Merged Images Influence the Effectiveness of Computer-Based Training in X-ray Screening?” *Security Technology (ICCST), 2011 IEEE International Carnahan Conference on Date of Conference*:18-21 Oct. 2011. pp.1 – 8.

[7] Stefan Michel, Adrian Schwaninger, “Human-Machine Interaction in X-ray Screening”, *Security Technology, 2009. 43rd Annual 2009 International Carnahan Conference on Date of Conference*:5-8 Oct. 2009. pp:13 – 19.

[8] Anton Bolfin, Adrian Schwaninger, “Selection and Pre-Employment Assessment in Aviation Security X-Ray Screening” *Security Technology, 2009. 43rd Annual 2009 International Carnahan Conference on Date of Conference*:5-8 Oct. 2009. pp:5 – 12.

[9] Diana Turcsany, Andre Mouton, Toby P. Breckon. “Improving Feature-Based Object Recognition for X-ray Baggage Security Screening Using Primed Visual Words”, *Industrial Technology (ICIT), 2013 IEEE International Conference on Date of Conference*:25-28 Feb. 2013. pp:1140 – 1145.

[10] Jianping Fan, Yuli Gao, Hangzai Luo and Ramesh Jain, “Mining Multilevel Image Semantics via Hierarchical Classification”, *IEEE Transactions on Multimedia*, vol. 10, NO. 2, Feb. 2008. pp:167-187.

[11] Rong Zhang, “Distinguishing Photorealistic Computer Graphics from Natural Images by Imaging Features and Visual Features”, *Electronics, Communications and Control (ICECC), 2011 International Conference on Date of Conference*: 9-11 Sept. 2011. PP:226 – 229.

[12] Jinhui Tang, “Semantic-Gap-Oriented Active Learning for Multilabel Image Annotation”, Image Processing, IEEE Transactions on (Volume:21 , Issue: 4) Date of Publication: April 2012. pp:2354 – 2360.

[13] Walther, D.B. “Using Confusion Matrices to Estimate Mutual Information between Two Categorical Measurements”, *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on Date of Conference*:22-24 June 2013. pp:220 – 224.

[14] G.Sengupta, T.A.Win, C.Messom, S.Demidenko and S.C.Mukhopadhyay, “Defect analysis of grit-blasted or spray printed surface using vision sensing technique”, Proceedings of Image and Vision Computing NZ, Nov. 26-28, 2003, Palmerston North, pp. 18-23.

[15] Jianjun Ma, Xianzhe Li. “Lab-on-a-Fiber Device for Trace Vapor TNT Explosive Detection: Comprehensive Performance Evaluation”, Lightwave Technology, Journal of (Volume:30 , Issue: 8) April 15, 2012. pp:1127 – 1133.

[16] Nicola Ivan Giannoccaro, A Robotic Arm to Sort Different Types of Ball Bearings from the Knowledge Discovered by Size Measurements of Image Regions and RFID Support, International Journal on Smart Sensing and Intelligent Systems, vol.7, no.2, pp. 674 – 700, 2014.

[17] Guohui Wu, Xingkun Li and Jiyang Dai, Improved Measure Algorithm Based on CoSaMP for Image Recovery vol.7, no.2, pp. 724 – 739, 2014.